Deterministic Image-to-Image Translation via Denoising Brownian Bridge Models with Dual Approximators
Bohan Xiao, Peiyong Wang, Qisheng He, Ming Dong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28232-28241
Image-to-Image (I2I) translation involves converting an im- age from one domain to another. Deterministic I2I transla- tion, such as in image super-resolution, extends this con- cept by guaranteeing that each input generates a consistent and predictable output, closely matching the ground truth (GT) with high fidelity. In this paper, we propose a denois- ing Brownian bridge model with dual approximators (Dual- approx Bridge), a novel generative model that exploits the Brownian bridge dynamics and two neural network-based approximators (one for forward and one for reverse pro- cess) to produce faithful output with negligible variance and high image quality in I2I translations. Our extensive exper- iments on benchmark datasets including image generation and super-resolution demonstrate the consistent and supe- rior performance of Dual-approx Bridge in terms of im- age quality and faithfulness to GT when compared to both stochastic and deterministic baselines. Project page and code: https://github.com/bohan95/dual-app-bridge
****

Towards Source-Free Machine Unlearning
Sk Miraj Ahmed, Umit Yigit Basaran, Dripta S. Raychaudhuri, Arindam Dutta, Rohit Kundu, Fahim Faisal Niloy, Basak Guler, Amit K. Roy-Chowdhury; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4948-4957
As machine learning become more pervasive and data privacy regulations evolve, the ability to remove private or copyrighted information from trained models is becoming an increasingly critical requirement. Existing unlearning methods often rely on the assumption of having access to the entire training dataset during the forgetting process. However, this assumption may not hold true in practical scenarios where the original training data may not be accessible, i.e., the source-free setting. To address this challenge, we focus on the source-free unlearning scenario, where an unlearning algorithm must be capable of removing specific data from a trained model without requiring access to the original training dataset. Building on recent work, we present a method that can estimate the Hessian of the unknown remaining training data, a crucial component required for efficient unlearning. Leveraging this estimation technique, our method enables efficient zero-shot unlearning while providing robust theoretical guarantees on the unlearning performance, while maintaining performance on the remaining data. Extensive experiments over a wide range of datasets verify the efficacy of our method.
****

Uni4D: Unifying Visual Foundation Models for 4D Modeling from a Single Video
David Yifan Yao, Albert J. Zhai, Shenlong Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1116-1126
This paper presents a unified approach to understanding dynamic scenes from casual videos. Large pretrained vision foundation models, such as vision-language, video depth prediction, motion tracking, and segmentation models, offer promising capabilities. However, training a single model for comprehensive 4D understanding remains challenging. We introduce Uni4D, a multi-stage optimization framework that harnesses multiple pretrained models to advance dynamic 3D modeling, including static/dynamic reconstruction, camera pose estimation, and dense 3D motion tracking. Our results show state-of-the-art performance in dynamic 4D modeling with superior visual quality. Notably, Uni4D requires no retraining or fine-tuning, highlighting the effectiveness of repurposing visual foundation models for 4D understanding. Code and more results are available at: https://davidyao99.github.io/uni4d.
****

DynScene: Scalable Generation of Dynamic Robotic Manipulation Scenes for Embodied AI
Sangmin Lee, Sungyong Park, Heewon Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12166-12175
Robotic manipulation in embodied AI critically depends on large-scale, high-qual

ity datasets that reflect realistic object interactions and physical dynamics. H
owever, existing data collection pipelines are often slow, expensive, and heavil
y reliant on manual efforts. We present DynScene, a diffusion-based framework fo
r generating dynamic robotic manipulation scenes directly from textual instructi
ons. Unlike prior methods that focus solely on static environments or isolated r
obot actions, DynScene decomposes the generation into two phases static scene sy
nthesis and action trajectory generation allowing fine-grained control and diver
sity. Our model enhances realism and physical feasibility through scene refineme
nt (layout sampling, quaternion quantization) and leverages residual action repr
esentation to enable action augmentation, generating multiple diverse trajectori
es from a single static configuration. Experiments show DynScene achieves 26.8x
faster generation, 1.84x higher accuracy, and 28% greater action diversity than
human-crafted data. Furthermore, agents trained with DynScene exhibit up to 19.4
% higher success rates across complex manipulation tasks. Our approach paves the
way for scalable, automated dataset generation in robot learning.
****

DiffLocks: Generating 3D Hair from a Single Image using Diffusion Models
Radu Alexandru Rosu, Keyu Wu, Yao Feng, Youyi Zheng, Michael J. Black; Proceedin
gs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1
0847-10857
We address the task of generating 3D hair geometry from a single image, which is
challenging due to the diversity of hairstyles and the lack of paired image-to-
3D hair data. Previous methods are primarily trained on synthetic data and cope
with the limited amount of such data by using low-dimensional intermediate repre
sentations, such as guide strands and scalp-level embeddings, that require post-
processing to decode, upsample, and add realism. These approaches fail to recons
truct detailed hair, struggle with curly hair, or are limited to handling only a
few hairstyles. To overcome these limitations, we propose DiffLocks, a novel fr
amework that enables detailed reconstruction of a wide variety of hairstyles dir
ectly from a single image. First, we address the lack of 3D hair data by automat
ing the creation of the largest synthetic hair dataset to date, containing 40K h
airstyles. Second, we leverage the synthetic hair dataset to learn an image-cond
itioned diffusion-transfomer model that generates accurate 3D strands from a sin
gle frontal image. By using a pretrained image backbone, our method generalizes
to in-the-wild images despite being trained only on synthetic data. Our diffusio
n model predicts a scalp texture map in which any point in the map contains the
latent code for an individual hair strand. These codes are directly decoded to 3
D strands without post-processing techniques. Representing individual strands, i
nstead of guide strands, enables the transformer to model the detailed spatial s
tructure of complex hairstyles. With this, DiffLocks can recover highly curled h
air, like afro hairstyles, from a single image for the first time. Data and code
is available at https://radualexandru.github.io/difflocks
****

Hyperbolic Category Discovery
Yuanpei Liu, Zhenqi He, Kai Han; Proceedings of the Computer Vision and Pattern
Recognition Conference (CVPR), 2025, pp. 9891-9900
Generalized Category Discovery (GCD) is an intriguing open-world problem that ha
s garnered increasing attention. Given a dataset that includes both labelled and
unlabelled images, GCD aims to categorize all images in the unlabelled subset,
regardless of whether they belong to known or unknown classes. In GCD, the commo
n practice typically involves applying a spherical projection operator at the en
d of the self-supervised pretrained backbone, operating within Euclidean or sphe
rical space. However, both of these spaces have been shown to be suboptimal for
encoding samples that possess hierarchical structures. In contrast, hyperbolic s
pace exhibits exponential volume growth relative to radius, making it inherently
strong at capturing the hierarchical structure of samples from both seen and un
seen categories. Therefore, we propose to tackle the category discovery challeng
e in the hyperbolic space. We introduce HypCD, a simple Hyperbolic framework for
learning hierarchy-aware representations and classifiers for generalized Catego
ry Discovery. HypCD first transforms the Euclidean embedding space of the backbo

ne network into hyperbolic space, facilitating subsequent representation and cla ssification learning by considering both hyperbolic distance and the angle betwe en samples. This approach is particularly helpful for knowledge transfer from kn own to unknown categories in GCD. We thoroughly evaluate HypCD on public GCD ben chmarks, by applying it to various baseline and state-of-the-art methods, consis tently achieving significant improvements.
****

The Language of Motion: Unifying Verbal and Non-verbal Language of 3D Human Moti on
Changan Chen, Juze Zhang, Shrinidhi K. Lakshmikanth, Yusu Fang, Ruizhi Shao, Gor don Wetzstein, Li Fei-Fei, Ehsan Adeli; Proceedings of the Computer Vision and P attern Recognition Conference (CVPR), 2025, pp. 6200-6211
Human communication is inherently multimodal, involving a combination of verbal and non-verbal cues such as speech, facial expressions, and body gestures. Model ing these behaviors is essential for understanding human interaction and for cre ating virtual characters that can communicate naturally in applications like gam es, films, and virtual reality. However, existing motion generation models are t ypically limited to specific input modalities--either speech, text, or motion da ta--and cannot fully leverage the diversity of available data. In this paper, we propose a novel framework that unifies verbal and non-verbal language using mul timodal language models for human motion understanding and generation. This mode l is flexible in taking text, speech, and motion or any combination of them as i nput. Coupled with our novel pre-training strategy, our model not only achieves state-of-the-art performance on co-speech gesture generation but also requires m uch less data for training. Our model also unlocks an array of novel tasks such as editable gesture generation and emotion prediction from motion. We believe un ifying the verbal and non-verbal language of human motion is essential for real- world applications, and language models offer a powerful approach to achieving t his goal.
****

CALICO: Part-Focused Semantic Co-Segmentation with Large Vision-Language Models
Kiet A. Nguyen, Adheesh Juvekar, Tianjiao Yu, Muntasir Wahed, Ismini Lourentzou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2 025, pp. 4550-4561
Recent advances in Large Vision-Language Models (LVLMs) have enabled general-pur pose vision tasks through visual instruction tuning. While existing LVLMs can ge nerate segmentation masks from text prompts for single images, they struggle wit h segmentation-grounded reasoning across images, especially at finer granularit ies such as object parts. In this paper, we introduce the new task of part-focuse d semantic co-segmentation, which involves identifying and segmenting common obj ects and their constituent common and unique parts across images. To address thi s task, we present CALICO, the first LVLM designed for multi-image part-level re asoning segmentation. CALICO features two key components, a novel Correspondence Extraction Module that identifies semantic part-level correspondences, and Corr espondence Adaptation Modules that embed this information into the LVLM to facil itate multi-image understanding in a parameter-efficient manner. To support trai ning and evaluation, we curate MixedParts, a large-scale multi-image segmentatio n dataset containing 2.4M samples across 44K images spanning diverse object an d part categories. Experimental results demonstrate that CALICO, with just 0.3% of its parameters finetuned, achieves strong performance on this challenging tas k.
****

Task Preference Optimization: Improving Multimodal Large Language Models with Vi sion Task Alignment
Ziang Yan, Zhilin Li, Yinan He, Chenting Wang, Kunchang Li, Xinhao Li, Xiangyu Z eng, Zilei Wang, Yali Wang, Yu Qiao, Limin Wang, Yi Wang; Proceedings of the Com puter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29880-29892
Current multimodal large language models (MLLMs) struggle with fine-grained or p recise understanding of visuals although they give comprehensive perception and reasoning in a spectrum of vision applications. Recent studies either develop to

ol-using or unify specific visual tasks into the autoregressive framework, often at the expense of overall multimodal performance. To address this issue and enhance MLLMs with visual tasks in a scalable fashion, we propose Task Preference Optimization (TPO), a novel method that utilizes differentiable task preferences derived from typical fine-grained visual tasks. TPO introduces learnable task tokens that establish connections between multiple task-specific heads and the MLLM. By leveraging rich visual labels during training, TPO significantly enhances the MLLM's multimodal capabilities and task-specific performance. Through multi-task co-training within TPO, we observe synergistic benefits that elevate individual task performance beyond what is achievable through single-task training methodologies. Our instantiation of this approach with VideoChat and LLaVA demonstrates an overall 14.6% improvement in multimodal performance compared to baseline models. Additionally, MLLM-TPO demonstrates robust zero-shot capabilities across various tasks, performing comparably to state-of-the-art supervised models.

****

Cross-modal Causal Relation Alignment for Video Question Grounding

Weixing Chen, Yang Liu, Binglin Chen, Jiandong Su, Yongsen Zheng, Liang Lin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24087-24096

Video question grounding (VideoQG) requires models to answer the questions and simultaneously infer the relevant video segments to support the answers. However, existing VideoQG methods usually suffer from spurious cross-modal correlations, leading to a failure to identify the dominant visual scenes that align with the intended question. Moreover, vision-language models exhibit unfaithful generalization performance and lack robustness on challenging downstream tasks such as VideoQG. In this work, we propose a novel VideoQG framework named Cross-modal Causal Relation Alignment (CRA), to eliminate spurious correlations and improve the causal consistency between question-answering and video temporal grounding. Our CRA involves three essential components: i) Gaussian Smoothing Grounding (GSG) module for estimating the time interval via cross-modal attention, which is de-noised by an adaptive Gaussian filter, ii) Cross-Modal Alignment (CMA) enhances the performance of weakly supervised VideoQG by leveraging bidirectional contrastive learning between estimated video segments and QA features, iii) Explicit Causal Intervention (ECI) module for multimodal deconfounding, which involves front-door intervention for vision and back-door intervention for language. Extensive experiments on two VideoQG datasets demonstrate the superiority of our CRA in discovering visually grounded content and achieving robust question reasoning. Codes are available at https://github.com/WissingChen/CRA-GQA.

****

Words or Vision: Do Vision-Language Models Have Blind Faith in Text?

Ailin Deng, Tri Cao, Zhirui Chen, Bryan Hooi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3867-3876

Vision-Language Models (VLMs) excel in integrating visual and textual information for vision-centric tasks, but their handling of inconsistencies between modalities is underexplored. We investigate VLMs' modality preferences when faced with visual data and varied textual inputs in vision-centered settings.By introducing textual variations to four vision-centric tasks and evaluating ten Vision-Language Models (VLMs), we discover a "blind faith in text" phenomenon: VLMs disproportionately trust textual data over visual data when inconsistencies arise, leading to significant performance drops under corrupted text and raising safety concerns.We analyze factors influencing this text bias, including instruction prompts, language model size, text relevance, token order, and the interplay between visual and textual certainty. While certain factors, such as scaling up the language model size, slightly mitigate text bias, others like token order can exacerbate it due to positional biases inherited from language models. To address this issue, we explore supervised fine-tuning with text augmentation and demonstrate its effectiveness in reducing text bias. Additionally, we provide a theoretical analysis suggesting that the blind faith in text phenomenon may stem from an imbalance of pure text and multi-modal data during training.Our findings highlight the need for balanced training and careful consideration of modality interactio

ns in VLMs to enhance their robustness and reliability in handling multi-modal data inconsistencies.
****
Diffusion Renderer: Neural Inverse and Forward Rendering with Video Diffusion Models
Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Chih-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, Zian Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26069-26080
Understanding and modeling lighting effects are fundamental tasks in computer vision and graphics. Classic physically-based rendering (PBR) accurately simulates the light transport, but relies on precise scene representations--explicit 3D geometry, high-quality material properties, and lighting conditions--that are often impractical to obtain in real-world scenarios. Therefore, we introduce Diffusion Renderer, a neural approach that addresses the dual problem of inverse and forward rendering within a holistic framework. Leveraging powerful video diffusion model priors, the inverse rendering model accurately estimates G-buffers from real-world videos, providing an interface for image editing tasks, and training data for the rendering model. Conversely, our rendering model generates photorealistic images from G-buffers without explicit light transport simulation. Specifically, we first train a video diffusion model for inverse rendering on synthetic data, which generalizes well to real-world videos and allows us to auto-label diverse real-world videos. We then co-train our rendering model using both synthetic and auto-labeled real-world data. Experiments demonstrate that Diffusion Renderer effectively approximates inverse and forwards rendering, consistently outperforming the state-of-the-art. Our model enables practical applications from a single video input--including relighting, material editing, and realistic object insertion.
****
Harnessing Frequency Spectrum Insights for Image Copyright Protection Against Diffusion Models
Zhenguang Liu, Chao Shuai, Shaojing Fan, Ziping Dong, Jinwu Hu, Zhongjie Ba, Kui Ren; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18653-18662
Diffusion models have achieved remarkable success in novel view synthesis, but their reliance on large, diverse, and often untraceable Web datasets has raised pressing concerns about image copyright protection. Current methods fall short in reliably identifying unauthorized image use, as they struggle to generalize across varied generation tasks and fail when the training dataset includes images from multiple sources with few identifiable (watermarked or poisoned) samples. In this paper, we present novel evidence that diffusion-generated images faithfully preserve the statistical properties of their training data, particularly reflected in their spectral features. Leveraging this insight, we introduce CoprGuard, a robust frequency domain watermarking framework to safeguard against unauthorized image usage in diffusion model training and fine-tuning. CoprGuard demonstrates remarkable effectiveness against a wide range of models, from naive diffusion models to sophisticated text-to-image models, and is robust even when watermarked images comprise a mere 1% of the training dataset. This robust and versatile approach empowers content owners to protect their intellectual property in the era of AI-driven image generation.
****
Learning to Detect Objects from Multi-Agent LiDAR Scans without Manual Labels
Qiming Xia, Wenkai Lin, Haoen Xiang, Xun Huang, Siheng Chen, Zhen Dong, Cheng Wang, Chenglu Wen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1418-1428
Unsupervised 3D object detection serves as an important solution for offline 3D object annotation. However, due to the data sparsity and limited views, the clustering-based label fitting in unsupervised object detection often generates low-quality pseudo-labels. Multi-agent collaborative dataset, which involves the sharing of complementary observations among agents, holds the potential to break th

rough this bottleneck. In this paper, we introduce a novel unsupervised method t hat learns to Detect Objects from Multi-Agent LiDAR scans, termed DOtA, without using labels from external. DOtA first uses the internally shared ego-pose and e go-shape of collaborative agents to initialize the detector, leveraging the gene ralization performance of neural networks to infer preliminary labels. Subsequen tly, DOtA uses the complementary observations between agents to perform multi-sc ale encoding on preliminary labels, then decodes high-quality and low-quality la bels. These labels are further used as prompts to guide a correct feature learni ng process, thereby enhancing the performance of the unsupervised object detecti on task. Extensive experiments on the V2V4Real and OPV2V datasets show that our DOtA outperforms state-of-the-art unsupervised 3D object detection methods. Addi tionally, we also validate the effectiveness of the DOtA labels under various co llaborative perception frameworks. The code is available at https://github.com/x muqimingxia/DOtA.

****

DeepLA-Net: Very Deep Local Aggregation Networks for Point Cloud Analysis

Ziyin Zeng, Mingyue Dong, Jian Zhou, Huan Qiu, Zhen Dong, Man Luo, Bijun Li; Pro ceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1330-1341

Due to the irregular and disordered data structure in 3D point clouds, prior wor ks have focused on designing more sophisticated local representation methods to capture these complex local patterns. However, the recognition performance has s aturated over the past few years, indicating that increasingly complex and redun dant designs no longer make improvements to local learning. This phenomenon prom pts us to diverge from the trend in 3D vision and instead pursue an alternative and successful solution: deeper neural networks. In this paper, we propose DeepL A-Net, a series of very deep networks for point cloud analysis. The key insight of our approach is to exploit a small but mighty local learning block, which red uces 10xfewer FLOPs, enabling the construction of very deep networks. Furthermor e, we design a training supervision strategy to ensure smooth gradient backpropa gation and optimization in very deep networks. We construct the DeepLA-Net famil y with a depth of up to 120 blocks --- at least 5xdeeper than recent methods --- trained on a single RTX 3090. An ensemble of the DeepLA-Net achieves state-of-t he-art performance on classification and segmentation tasks of S3DIS Area5 (+2.2 % mIoU), ScanNet test set (+1.6% mIoU), ScanObjectNN (+2.1% OA), and ShapeNetPar t (+0.9% cls.mIoU).

****

Multi-Layer Visual Feature Fusion in Multimodal LLMs: Methods, Analysis, and Bes t Practices

Junyan Lin, Haoran Chen, Yue Fan, Yingqi Fan, Xin Jin, Hui Su, Jinlan Fu, Xiaoyu Shen; Proceedings of the Computer Vision and Pattern Recognition Conference (CV PR), 2025, pp. 4156-4166

Multimodal Large Language Models (MLLMs) have made significant advancements in r ecent years, with visual features playing an increasingly critical role in enhan cing model performance. However, the integration of multi-layer visual features in MLLMs remains underexplored, particularly with regard to optimal layer select ion and fusion strategies. Existing methods often rely on arbitrary design choic es, leading to suboptimal outcomes. In this paper, we systematically investigate two core aspects of multi-layer visual feature fusion: (1) selecting the most e ffective visual layers and (2) identifying the best fusion approach with the lan guage model. Our experiments reveal that while combining visual features from mu ltiple stages improves generalization, incorporating additional features from th e same stage typically leads to diminished performance. Furthermore, we find tha t direct fusion of multi-layer visual features at the input stage consistently y ields superior and more stable performance across various configurations.

****

APHQ-ViT: Post-Training Quantization with Average Perturbation Hessian Based Rec onstruction for Vision Transformers

Zhuguanyu Wu, Jiayi Zhang, Jiaxin Chen, Jinyang Guo, Di Huang, Yunhong Wang; Pro ceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025,

pp. 9686-9695
Vision Transformers (ViTs) have become one of the most commonly used backbones for vision tasks. Despite their remarkable performance, they often suffer significant accuracy drop when quantized for practical deployment, particularly by post-training quantization (PTQ) under ultra-low bits. Recently, reconstruction-based PTQ methods have shown promising performance in quantizing Convolutional Neural Networks (CNNs). However, they fail when applied to ViTs, primarily due to the inaccurate estimation of output importance and the substantial accuracy degradation in quantizing post-GELU activations. To address these issues, we propose APHQ-ViT, a novel PTQ approach based on importance estimation with Average Perturbation Hessian (APH). Specifically, we first thoroughly analyze the current approximation approaches with Hessian loss, and propose an improved average perturbation Hessian loss. To deal with the quantization of the post-GELU activations, we design an MLP-Reconstruction (MR) method by replacing the GELU function in MLP with ReLU and reconstructing it by the APH loss on a small unlabeled calibration set. Extensive experiments demonstrate that APHQ-ViT using linear quantizers outperforms existing PTQ methods by substantial margins in 3-bit and 4-bit across different vision tasks. The source code is available at https://github.com/GoatWu/APHQ-ViT.

****

AdaptCMVC: Robust Adaption to Incremental Views in Continual Multi-view Clustering

Jing Wang, Songhe Feng, Kristoffer Knutsen Wickstrøm, Michael C. Kampffmeyer; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10285-10294

Most Multi-view Clustering approaches assume that all views are available for clustering. However, this assumption is often unrealistic as views are incrementally accumulated over time, leading to a need for continual multi-view clustering (CMVC) methods. Current approaches to CMVC leverage late fusion-based approaches, where a new model is typically learned individually for each view to obtain the corresponding partition matrix, and then used to update a consensus matrix via a moving average. These approaches are prone to view-specific noise and struggle to adapt to large gaps between different views. To address these shortcomings, we reconsider CMVC from the perspective of domain adaptation and propose AdaptCMVC, which learns how to incrementally accumulate knowledge of new views as they become available and prevents catastrophic forgetting. Specifically, a self-training framework is introduced to extend the model to new views, particularly designed to be robust to view-specific noise. Further, to combat catastrophic forgetting, a structure alignment mechanism is proposed to enable the model to explore the global group structure across multiple views. Experiments on several multi-view benchmarks demonstrate the effectiveness of our proposed method on the CMVC task. The code is available at: AdaptCMVC.

****

Omni-Scene: Omni-Gaussian Representation for Ego-Centric Sparse-View Scene Reconstruction

Dongxu Wei, Zhiqi Li, Peidong Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22317-22327

Prior works employing pixel-based Gaussian representation have demonstrated efficacy in feed-forward sparse-view reconstruction. However, such representation necessitates cross-view overlap for accurate depth estimation, and is challenged by object occlusions and frustum truncations. As a result, these methods require scene-centric data acquisition to maintain cross-view overlap and complete scene visibility to circumvent occlusions and truncations, which limits their applicability to scene-centric reconstruction. In contrast, in autonomous driving scenarios, a more practical paradigm is ego-centric reconstruction, which is characterized by minimal cross-view overlap and frequent occlusions and truncations. The limitations of pixel-based representation thus hinder the utility of prior works in this task. In light of this, this paper conducts an in-depth analysis of different representations, and introduces Omni-Gaussian representation with tailored network design to complement their strengths and mitigate their drawbacks. Ex

periments show that our method significantly surpasses state-of-the-art methods, pixelSplat and MVSplat, in ego-centric reconstruction, and achieves comparable performance to prior works in scene-centric reconstruction. Our code is availabl e at https://github.com/WU-CVGL/Omni-Scene.

****

3DTopia-XL: Scaling High-quality 3D Asset Generation via Primitive Diffusion

Zhaoxi Chen, Jiaxiang Tang, Yuhao Dong, Ziang Cao, Fangzhou Hong, Yushi Lan, Ten gfei Wang, Haozhe Xie, Tong Wu, Shunsuke Saito, Liang Pan, Dahua Lin, Ziwei Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2 025, pp. 26576-26586

The increasing demand for high-quality 3D assets across various industries neces sitates efficient and automated 3D content creation. Despite recent advancements in 3D generative models, existing methods still face challenges with optimizati on speed, geometric fidelity, and the lack of assets for physically based render ing (PBR). In this paper, we introduce 3DTopia-XL, a scalable native 3D generati ve model designed to overcome these limitations. 3DTopia-XL leverages a novel pr imitive-based 3D representation, PrimX, which encodes detailed shape, albedo, an d material field into a compact tensorial format, facilitating the modeling of h igh-resolution geometry with PBR assets. On top of the novel representation, we propose a generative framework based on Diffusion Transformer (DiT), which compr ises 1) Primitive Patch Compression, 2) and Latent Primitive Diffusion. 3DTopia-XL learns to generate high-quality 3D assets from textual or visual inputs. Exte nsive qualitative and quantitative experiments are conducted to demonstrate that 3DTopia-XL significantly outperforms existing methods in generating high-qualit y 3D assets with fine-grained textures and materials, efficiently bridging the q uality gap between generative models and real-world applications.

****

UA-Pose: Uncertainty-Aware 6D Object Pose Estimation and Online Object Completio n with Partial References

Ming-Feng Li, Xin Yang, Fu-En Wang, Hritam Basak, Yuyin Sun, Shreekant Gayaka, M in Sun, Cheng-Hao Kuo; Proceedings of the Computer Vision and Pattern Recognitio n Conference (CVPR), 2025, pp. 1180-1189

6D object pose estimation has shown strong generalizability to novel objects. Ho wever, existing methods often require either a complete, well-reconstructed 3D m odel or numerous reference images that fully cover the object. Estimating 6D pos es from partial references, which capture only fragments of an object's appearan ce and geometry, remains challenging. To address this, we propose UA-Pose, an un certainty-aware approach for 6D object pose estimation and online object complet ion specifically designed for partial references. We assume access to either (1) a limited set of RGBD images with known poses or (2) a single 2D image. For the first case, we initialize a partial object 3D model based on the provided image s and poses, while for the second, we use image-to-3D techniques to generate an initial object 3D model. Our method integrates uncertainty into the incomplete 3 D model, distinguishing between seen and unseen regions. This uncertainty enable s confidence assessment in pose estimation and guides an uncertainty-aware sampl ing strategy for online object completion, enhancing robustness in pose estimati on accuracy and improving object completeness. We evaluate our method on the YCB -Video, YCBInEOAT, and HO3D datasets, including RGBD sequences of YCB objects ma nipulated by robots and human hands. Experimental results demonstrate significan t performance improvements over existing methods, particularly when object obser vations are incomplete or partially captured.

****

Missing Target-Relevant Information Prediction with World Model for Accurate Zer o-Shot Composed Image Retrieval

Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Gaopeng Gou, Qi Wu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 20 25, pp. 24785-24795

Zero-Shot Composed Image Retrieval (ZS-CIR) involves diverse tasks with a broad range of visual content manipulation intent across domain, scene, object, and at tribute. The key challenge for ZS-CIR tasks is to modify a reference image accor

ding to manipulation text to accurately retrieve a target image, especially when the reference image is missing essential target content. In this paper, we propose a novel prediction-based mapping network, named PrediCIR, to adaptively predict the missing target visual content in reference images in the latent space before mapping for accurate ZS-CIR. Specifically, a world view generation module first constructs a source view by omitting certain visual content of a target view, coupled with an action that includes the manipulation intent derived from existing image-caption pairs. Then, a target content prediction module trains a world model as a predictor to adaptively predict the missing visual information guided by user intention in manipulating text at the latent space. The two modules map an image with the predicted relevant information to a pseudo-word token without extra supervision. Our model shows strong generalization ability on six ZS-CIR tasks. It obtains consistent and significant performance boosts ranging from 1.73% to 4.45% over the best methods and achieves new state-of-the-art results on ZS-CIR. Our code is available at https://github.com/Pter61/predicir.
****

Binarized Mamba-Transformer for Lightweight Quad Bayer HybridEVS Demosaicing
Shiyang Zhou, Haijin Zeng, Yunfan Lu, Tong Shao, Ke Tang, Yongyong Chen, Jie Liu, Jingyong Su; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8817-8827
Quad Bayer demosaicing is the central challenge for enabling the widespread application of Hybrid Event-based Vision Sensors (HybridEVS). Although existing learning-based methods that leverage long-range dependency modeling have achieved promising results, their complexity severely limits deployment on mobile devices for real-world applications. To address these limitations, we propose a lightweight Mamba-based binary neural network designed for efficient and high-performing demosaicing of HybridEVS RAW images. First, to effectively capture both global and local dependencies, we introduce a hybrid Binarized Mamba-Transformer architecture that combines the strengths of the Mamba and Swin Transformer architectures. Next, to significantly reduce computational complexity, we propose a binarized Mamba (Bi-Mamba), which binarizes all projections while retaining the core Selective Scan in full precision. Bi-Mamba also incorporates additional global visual information to enhance global context and mitigate precision loss. We conduct quantitative and qualitative experiments to demonstrate the effectiveness of BMTNet in both performance and computational efficiency, providing a lightweight demosaicing solution suited for real-world edge devices. Our codes and models are available at https://github.com/Clausy9/BMTNet.
****

DiffSensei: Bridging Multi-Modal LLMs and Diffusion Models for Customized Manga Generation
Jianzong Wu, Chao Tang, Jingbo Wang, Yanhong Zeng, Xiangtai Li, Yunhai Tong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28684-28693
Story visualization, the task of creating visual narratives from textual descriptions, has seen progress with text-to-image generation models. However, these models often lack effective control over character appearances and interactions, particularly in multi-character scenes. To address these limitations, we propose a new task: customized manga generation and introduce DiffSensei, an innovative framework specifically designed for generating manga with dynamic multi-character control. DiffSensei integrates a diffusion-based image generator with a multimodal large language model (MLLM) that acts as a text-compatible identity adapter. Our approach employs masked cross-attention to seamlessly incorporate character features, enabling precise layout control without direct pixel transfer. Additionally, the MLLM-based adapter adjusts character features to align with panel-specific text cues, allowing flexible adjustments in character expressions, poses, and actions. We also introduce MangaZero, a large-scale dataset tailored to this task, containing 43,264 manga pages and 427,147 annotated panels, supporting the visualization of varied character interactions and movements across sequential frames. Extensive experiments demonstrate that DiffSensei outperforms existing models, marking a significant advancement in manga generation by enabling text

-adaptable character customization. The code, model, and dataset will be open-so
urced to the community.
****

Narrating the Video: Boosting Text-Video Retrieval via Comprehensive Utilization
 of Frame-Level Captions
Chan Hur, Jeong-hun Hong, Dong-hun Lee, Dabin Kang, Semin Myeong, Sang-hyo Park,
 Hyeyoung Park; Proceedings of the Computer Vision and Pattern Recognition Confe
rence (CVPR), 2025, pp. 24077-24086
In recent text-video retrieval, the use of additional captions from vision-langu
age models has shown promising effects on the performance. However, existing mod
els using additional captions often have struggled to capture the rich semantics
, including temporal changes, inherent in the video. In addition, incorrect info
rmation caused by generative models can lead to inaccurate retrieval. To address
 these issues, we propose a new framework, Narrating the Video (NarVid), which s
trategically leverages the comprehensive information available from frame-level
captions, the narration. The proposed NarVid exploits narration in multiple ways
: 1) feature enhancement through cross-modal interactions between narration and
video, 2) query-aware adaptive filtering to suppress irrelevant or incorrect inf
ormation, 3) dual-modal matching score by adding query-video similarity and quer
y-narration similarity, and 4) hard-negative loss to learn discriminative featur
es from multiple perspectives using the two similarities from different views. E
xperimental results demonstrate that NarVid achieves state-of-the-art performanc
e on various benchmark datasets.
****

IDEA-Bench: How Far are Generative Models from Professional Designing?
Chen Liang, Lianghua Huang, Jingwu Fang, Huanzhang Dou, Wei Wang, Zhi-Fan Wu, Yu
peng Shi, Junge Zhang, Xin Zhao, Yu Liu; Proceedings of the Computer Vision and
Pattern Recognition Conference (CVPR), 2025, pp. 18541-18551
Recent advancements in image generation models enable the creation of high-quali
ty images and targeted modifications based on textual instructions. Some models
even support multimodal complex guidance and demonstrate robust task generalizat
ion capabilities. However, they still fall short of meeting the nuanced, profess
ional demands of designers. To bridge this gap, we introduce IDEA-Bench, a compr
ehensive benchmark designed to advance image generation models toward applicatio
ns with robust task generalization. IDEA-Bench comprises 100 professional image
generation tasks and 275 specific cases, categorized into five major types based
 on the current capabilities of existing models. Furthermore, we provide a repre
sentative subset of 18 tasks with enhanced evaluation criteria to facilitate mor
e nuanced and reliable evaluations using Multimodal Large Language Models (MLLMs
). By assessing models' ability to comprehend and execute novel, complex tasks,
IDEA-Bench paves the way toward the development of generative models with autono
mous and versatile visual generation capabilities.
****

Interpretable Image Classification via Non-parametric Part Prototype Learning
Zhijie Zhu, Lei Fan, Maurice Pagnucco, Yang Song; Proceedings of the Computer Vi
sion and Pattern Recognition Conference (CVPR), 2025, pp. 9762-9771
Classifying images with an interpretable decision-making process is a long-stand
ing problem in computer vision. In recent years, Prototypical Part Networks has
gained traction as an approach for self-explainable neural networks, due to thei
r ability to mimic human visual reasoning by providing explanations based on pro
totypical object parts. However, the quality of the explanations generated by th
ese methods leaves room for improvement, as the prototypes usually focus on repe
titive and redundant concepts. Leveraging recent advances in prototype learning,
 we present a framework for part-based interpretable image classification that l
earns a set of semantically distinctive object parts for each class, and provide
s diverse and comprehensive explanations. The core of our method is to learn the
 part-prototypes in a non-parametric fashion, through clustering deep features e
xtracted from foundation vision models that encode robust semantic information.
To quantitatively evaluate the quality of explanations provided by ProtoPNets, w
e introduce Distinctiveness Score and Comprehensiveness Score. Through evaluatio

n on CUB-200-2011, Stanford Cars and Stanford Dogs datasets, we show that our fr amework compares favourably against existing ProtoPNets while achieving better i nterpretability.
****

PhD: A ChatGPT-Prompted Visual Hallucination Evaluation Dataset
Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhan hui Kang, Xirong Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19857-19866
Multimodal Large Language Models (MLLMs) hallucinate, resulting in an emerging t opic of visual hallucination evaluation (VHE). This paper contributes a ChatGPT-Prompted visual hallucination evaluation Dataset (PhD) for objective VHE at a la rge scale. The essence of VHE is to ask an MLLM questions about specific images to assess its susceptibility to hallucination. Depending on what to ask (objects , attributes, sentiment, etc.) and how the questions are asked, we structure PhD along two dimensions, i.e. task and mode. Five visual recognition tasks, rangin g from low-level (object / attribute recognition) to middle-level (sentiment / p osition recognition and counting), are considered. Besides a normal visual QA mo de, which we term PhD-base, PhD also asks questions with specious context (PhD-s ec) or with incorrect context (PhD-icc), or with AI-generated counter common sen se images (PhD-ccs). We construct PhD by a ChatGPT-assisted semi-automated pipel ine, encompassing four pivotal modules: task-specific hallucinatory item (hitem) selection, hitem-embedded question generation, specious / incorrect context gen eration, and counter-common-sense (CCS) image generation. With over 14k daily im ages, 750 CCS images and 102k VQA triplets in total, PhD reveals considerable va riability in MLLMs' performance across various modes and tasks, offering valuabl e insights into the nature of hallucination. As such, PhD stands as a potent too l not only for VHE but may also play a significant role in the refinement of MLL Ms.
****

CARL: A Framework for Equivariant Image Registration
Hastings Greer, Lin Tian, François-Xavier Vialard, Roland Kwitt, Raul San Jose E stepar, Marc Niethammer; Proceedings of the Computer Vision and Pattern Recognit ion Conference (CVPR), 2025, pp. 26014-26023
Image registration estimates spatial correspondences between image pairs. These estimates are typically obtained via numerical optimization or regression by a d eep network. A desirable property is that a correspondence estimate (e.g., the t rue oracle correspondence) for an image pair is maintained under deformations of the input images. Formally, the estimator should be equivariant to a desired cl ass of image transformations. In this work, we present careful analyses of equiv ariance properties in the context of multi-step deep registration networks. Base d on these analyses we 1) introduce the notions of [U,U] equivariance (network e quivariance to the same deformations of the input images) and [W,U] equivariance (where input images can undergo different deformations); we 2) show that in a s uitable multi-step registration setup it is sufficient for overall [W,U] equivar iance if the first step has [W,U] equivariance and all others have [U,U] equivar iance; we 3) show that common displacement-predicting networks only exhibit [U,U] equivariance to translations instead of the more powerful [W,U] equivariance; and we 4) show how to achieve multi-step [W,U] equivariance via a coordinate-att ention mechanism combined with displacement-predicting networks. Our approach ob tains excellent practical performance for 3D abdomen, lung, and brain medical im age registration. We match or outperform state-of-the-art (SOTA) registration ap proaches on all the datasets with a particularly strong performance for the chal lening abdomen registration.
****

ClimbingCap: Multi-Modal Dataset and Method for Rock Climbing in World Coordinat e
Ming Yan, Xincheng Lin, Yuhua Luo, Shuqi Fan, Yudi Dai, Qixin Zhong, Lincai Zhon g, Yuexin Ma, Lan Xu, Chenglu Wen, Siqi Shen, Cheng Wang; Proceedings of the Com puter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12312-12323
Human Motion Recovery (HMR) research mainly focuses on ground-based motions such

as running. The study on capturing climbing motion, an off-ground motion, is sparse. This is partly due to the limited availability of climbing motion datasets, especially large-scale and challenging 3D labeled datasets. To address the insufficiency of climbing motion datasets, we collect AscendMotion, a large-scale well-annotated, and challenging climbing motion dataset. It consists of 412k RGB, LiDAR frames, and IMU measurements, which includes the challenging climbing motions of 22 professional climbing coaches across 12 different rocks. Capturing the climbing motions is challenging as it requires precise recovery of not only the complex pose but also the global position of climbers. Although multiple global HMR methods have been proposed, they cannot faithfully capture climbing motions. To address the limitations of HMR methods for climbing, we propose ClimbingCap, a motion recovery method that reconstructs continuous 3D human climbing motion in a global coordinate system. One key insight is to use the RGB and the LiDAR modalities to separately reconstruct motions in camera coordinates and global coordinates and optimize them jointly. We demonstrate the quality of the AscendMotion dataset and present promising results from ClimbingCap. The AscendMotion dataset and the source code of ClimbingCap will be released publicly to the research community.

****

DAGSM: Disentangled Avatar Generation with GS-enhanced Mesh

Jingyu Zhuang, Di Kang, Linchao Bao, Liang Lin, Guanbin Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 292-303

Text-driven avatar generation has gained significant attention owing to its convenience. However, existing methods typically model the human body with all garments as a single 3D model, limiting its usability, such as clothing replacement, and reducing user control over the generation process. To overcome the limitations above, we propose DAGSM, a novel pipeline that generates disentangled human bodies and garments from the given text prompts. Specifically, we model each part (e.g., body, upper/lower clothes) of the clothed human as one GS-enhanced mesh (GSM), which is a traditional mesh attached with 2D Gaussians to better handle complicated textures (e.g., woolen, translucent clothes) and produce realistic cloth animations. During the generation, we first create the unclothed body, followed by a sequence of individual cloth generation based on the body, where we introduce a semantic-based algorithm to achieve better human-cloth and garment-garment separation. To improve texture quality, we propose a view-consistent texture refinement module, including a cross-view attention mechanism for texture style consistency and an incident-angle-weighted denoising (IAW-DE) strategy to update the appearance. Extensive experiments have demonstrated that DAGSM generates high-quality disentangled avatars, supports clothing replacement and realistic animation, and outperforms the baselines in visual quality.

****

Estimating Body and Hand Motion in an Ego-sensed World

Brent Yi, Vickie Ye, Maya Zheng, Yunqi Li, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, Angjoo Kanazawa; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7072-7084

We present EgoAllo, a system for human motion estimation from a head-mounted device. Using only egocentric SLAM poses and images, EgoAllo guides sampling from a conditional diffusion model to estimate 3D body pose, height, and hand parameters that capture a device wearer's actions in the allocentric coordinate frame of the scene. To achieve this, our key insight is in representation: we propose spatial and temporal invariance criteria for improving model performance, from which we derive a head motion conditioning parameterization that improves estimation by up to 18%. We also show how the bodies estimated by our system can improve hand estimation: the resulting kinematic and temporal constraints can reduce world-frame errors in single-frame estimates by 40%.

****

A Bias-Free Training Paradigm for More General AI-generated Image Detection

Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, Luisa Verdoliva; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18685-18694

Successful forensic detectors can produce excellent results in supervised learni
ng benchmarks but struggle to transfer to real-world applications. We believe th
is limitation is largely due to inadequate training data quality. While most res
earch focuses on developing new algorithms, less attention is given to training
 data selection, despite evidence that performance can be strongly impacted by sp
urious correlations such as content, format, or resolution. A well-designed fore
nsic detector should detect generator specific artifacts rather than reflect dat
a biases. To this end, we propose B-Free, a bias-free training paradigm, where f
ake images are generated from real ones using the conditioning procedure of stab
le diffusion models. This ensures semantic alignment between real and fake image
s, allowing any differences to stem solely from the subtle artifacts introduced
 by AI generation. Through content-based augmentation, we show significant improv
ements in both generalization and robustness over state-of-the-art detectors and
 more calibrated results across 27 different generative models, including recent
 releases, like FLUX and Stable Diffusion 3.5. Our findings emphasize the import
ance of a careful dataset design, highlighting the need for further research on
 this topic. Code and data are publicly available at https://grip-unina.github.io
/B-Free/.

****

FALCON: Fairness Learning via Contrastive Attention Approach to Continual Semant
ic Scene Understanding

Thanh-Dat Truong, Utsav Prabhu, Bhiksha Raj, Jackson Cothren, Khoa Luu; Proceedi
ngs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp.
 15065-15075

Continual Learning in semantic scene segmentation aims to continually learn new
 unseen classes in dynamic environments while maintaining previously learned know
ledge. Prior studies focused on modeling the catastrophic forgetting and backgro
und shift challenges in continual learning. However, fairness, another major cha
llenge that causes unfair predictions leading to low performance among major and
 minor classes, still needs to be well addressed. In addition, prior methods hav
e yet to model the unknown classes well, thus resulting in producing non-discrim
inative features among unknown classes. This work presents a novel Fairness Lear
ning via Contrastive Attention Approach to continual learning in semantic scene
 understanding. In particular, we first introduce a new Fairness Contrastive Clus
tering loss to address the problems of catastrophic forgetting and fairness. The
n, we propose an attention-based visual grammar approach to effectively model th
e background shift problem and unknown classes, producing better feature represe
ntations for different unknown classes. Through our experiments, our proposed ap
proach achieves State-of-the-Art (SoTA) performance on different continual learn
ing benchmarks, i.e., ADE20K, Cityscapes, and Pascal VOC. It promotes the fairne
ss of the continual semantic segmentation model.

****

Certified Human Trajectory Prediction

Mohammadhossein Bahari, Saeed Saadatnejad, Amirhossein Askari Farsangi, Seyed-Mo
hsen Moosavi-Dezfooli, Alexandre Alahi; Proceedings of the Computer Vision and P
attern Recognition Conference (CVPR), 2025, pp. 12301-12311

Predicting human trajectories is essential for the safe operation of autonomous
 vehicles, yet current data-driven models often lack robustness in case of noisy
 inputs such as adversarial examples or imperfect observations. Although some tra
jectory prediction methods have been developed to provide empirical robustness,
 these methods are heuristic and do not offer guaranteed robustness. In this work
, we propose a certification approach tailored for trajectory prediction that pr
ovides guaranteed robustness. To this end, we address the unique challenges asso
ciated with trajectory prediction, such as unbounded outputs and multi-modality.
 To mitigate the inherent performance drop through certification, we propose a d
iffusion-based trajectory denoiser and integrate it into our method. Moreover, w
e introduce new certified performance metrics to reliably measure the trajectory
 prediction performance. Through comprehensive experiments, we demonstrate the a
ccuracy and robustness of the certified predictors and highlight their advantage
s over the non-certified ones. The code is available online: https://s-attack.gi

thub.io/
****
Evaluating Vision-Language Models as Evaluators in Path Planning
Mohamed Aghzal, Xiang Yue, Erion Plaku, Ziyu Yao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6886-6897
Despite their promise to perform complex reasoning, large language models (LLMs) have been shown to have limited effectiveness in end-to-end planning. This has inspired an intriguing question: if these models cannot plan well, can they still contribute to the planning framework as a helpful plan evaluator? In this work, we generalize this question to consider LLMs augmented with visual understanding, i.e., Vision-Language Models (VLMs). We introduce **PathEval**, a novel benchmark evaluating VLMs as plan evaluators in complex path-planning scenarios. Succeeding in the benchmark requires a VLM to be able to abstract traits of optimal paths from the scenario description, demonstrate precise low-level perception on each path, and integrate this information to decide the better path. Our analysis of state-of-the-art VLMs reveals that these models face significant challenges on the benchmark. We observe that the VLMs can precisely abstract given scenarios to identify the desired traits and exhibit mixed performance in integrating the provided information. Yet, their vision component presents a critical bottleneck, with models struggling to perceive low-level details about a path. Our experimental results show that this issue cannot be trivially addressed via end-to-end fine-tuning; rather, task-specific discriminative adaptation of these vision encoders is needed for these VLMs to become effective path evaluators.
****
Free on the Fly: Enhancing Flexibility in Test-Time Adaptation with Online EM
Qiyuan Dai, Sibei Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9538-9548
Vision-Language Models (VLMs) have become prominent in open-world image recognition for their strong generalization abilities. Yet, their effectiveness in practical applications is compromised by domain shifts and distributional changes, especially when test data distributions diverge from training data. Therefore, the paradigm of test-time adaptation (TTA) has emerged, enabling the use of online off-the-shelf data at test time, supporting independent sample predictions, and eliminating reliance on test annotations. Traditional TTA methods, however, often rely on costly training or optimization processes, or make unrealistic assumptions about accessing or storing historical training and test data.Instead, this study proposes FreeTTA, a training-free and universally available method that makes no assumptions, to enhance the flexibility of TTA. More importantly, FreeTTA is the first to explicitly model the test data distribution, enabling the use of intrinsic relationships among test samples to enhance predictions of individual samples without simultaneous access--a direction not previously explored. FreeTTA achieves these advantages by introducing an online EM algorithm that utilizes zero-shot predictions from VLMs as priors to iteratively compute the posterior probabilities of each online test sample and update parameters. Experiments demonstrate that FreeTTA achieves stable and significant improvements compared to state-of-the-art methods across 15 datasets in both cross-domain and out-of-distribution settings.
****
Transformers without Normalization
Jiachen Zhu, Xinlei Chen, Kaiming He, Yann LeCun, Zhuang Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14901-14911
Normalization layers are ubiquitous in modern neural networks and have long been considered essential. This work demonstrates that Transformers without normalization can achieve the same or better performance using a remarkably simple technique. We introduce Dynamic Tanh (DyT), an element-wise operation DyT(x)=tanh(ax), as a drop-in replacement for normalization layers in Transformers. DyT is inspired by the observation that layer normalization in Transformers often produces tanh-like, S-shaped input-output mappings. By incorporating DyT, Transformers without normalization can match or exceed the performance of their normalized coun

terparts, mostly without hyperparameter tuning. We validate the effectiveness of Transformers with DyT across diverse settings, ranging from recognition to generation, supervised to self-supervised learning, and computer vision to language models. These findings challenge the conventional understanding that normalization layers are indispensable in modern neural networks, and offer new insights into their role in deep networks.

****

SGC-Net: Stratified Granular Comparison Network for Open-Vocabulary HOI Detection

Xin Lin, Chong Shi, Zuopeng Yang, Haojin Tang, Zhili Zhou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4539-4549

Recent open-vocabulary human-object interaction (OV-HOI) detection methods primarily rely on large language model (LLM) for generating auxiliary descriptions and leverage knowledge distilled from CLIP to detect unseen interaction categories. Despite their effectiveness, these methods face two challenges: (1) feature granularity deficiency, due to reliance on last layer visual features for text alignment, leading to the neglect of crucial object-level details from intermediate layers; (2) semantic similarity confusion, resulting from CLIP's inherent biases toward certain classes, while LLM-generated descriptions based solely on labels fail to adequately capture inter-class similarities. To address these challenges, we propose a stratified granular comparison network. First, we introduce a granularity sensing alignment module that aggregates global semantic features with local details, refining interaction representations and ensuring robust alignment between intermediate visual features and text embeddings. Second, we develop a hierarchical group comparison module that recursively compares and groups classes using LLMs, generating fine-grained and discriminative descriptions for each interaction category. Experimental results on two widely-used benchmark datasets, SWIG-HOI and HICO-DET, demonstrate that our method achieves state-of-the-art results in OV-HOI detection. Codes is available at https://github.com/Phil0212/SGC-Net.

****

Galaxy Walker: Geometry-aware VLMs For Galaxy-scale Understanding

Tianyu Chen, Xingcheng Fu, Yisen Gao, Haodong Qian, Yuecen Wei, Kun Yan, Haoyi Zhou, Jianxin Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4112-4121

Modern vision-language models (VLMs) develop patch embedding and convolution backbone within vector space, especially Euclidean ones, at the very founding. When expanding VLMs to a galaxy-scale for understanding astronomical phenomena, the integration of spherical space for planetary orbits and hyperbolic spaces for black holes raises two formidable challenges. a) The current pre-training model is confined to Euclidean space rather than a comprehensive geometric embedding. b) The predominant architecture lacks suitable backbones for anisotropic physical geometries. In this paper, we introduced Galaxy-Walker, a geometry-aware VLM, for the universe-level vision understanding tasks. We proposed the geometry prompt that generates geometry tokens by random walks across diverse spaces on a multi-scale physical graph, along with a geometry adapter that compresses and reshapes the space anisotropy in a mixture-of-experts manner. Extensive experiments demonstrate the effectiveness of our approach, with Galaxy-Walker achieving state-of-the-art performance in both galaxy property estimation (R2 scores up to 0.91) and morphology classification tasks (up to +0.17 F1 improvement in challenging features), significantly outperforming both domain-specific models and general-purpose VLMs.

****

HiPART: Hierarchical Pose AutoRegressive Transformer for Occluded 3D Human Pose Estimation

Hongwei Zheng, Han Li, Wenrui Dai, Ziyang Zheng, Chenglin Li, Junni Zou, Hongkai Xiong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16807-16817

Existing 2D-to-3D human pose estimation (HPE) methods struggle with the occlusion issue by enriching information like temporal and visual cues in the lifting st

age. In this paper, we argue that these methods ignore the limitation of the spa rse skeleton 2D input representation, which fundamentally restricts the 2D-to-3D lifting and worsens the occlusion issue. To address these, we propose a novel t wo-stage generative densification method, named Hierarchical Pose AutoRegressive Transformer (HiPART), to generate hierarchical 2D dense poses from the origina l sparse 2D pose. Specifically, we first develop a multi-scale skeleton tokeniza tion module to quantize the highly dense 2D pose into hierarchical tokens and pr opose a skeleton-aware alignment to strengthen token connections. We then develo p a hierarchical autoregressive modeling scheme for hierarchical 2D pose generat ion. With generated hierarchical poses as inputs for 2D-to-3D lifting, the propo sed method shows strong robustness in occluded scenarios and achieves state-of-t he-art performance on the single-frame-based 3D HPE. Moreover, it outperforms nu merous multi-frame methods while reducing parameter and computational complexity and can also complement them to further enhance performance and robustness.
****

SnowMaster: Comprehensive Real-world Image Desnowing via MLLM with Multi-Model F eedback Optimization

Jianyu Lai, Sixiang Chen, Yunlong Lin, Tian Ye, Yun Liu, Song Fei, Zhaohu Xing, Hongtao Wu, Weiming Wang, Lei Zhu; Proceedings of the Computer Vision and Patter n Recognition Conference (CVPR), 2025, pp. 4302-4312

Snowfall presents significant challenges for visual data processing, necessitati ng specialized desnowing algorithms. However, existing models often fail to gene ralize effectively due to their heavy reliance on synthetic datasets. Furthermor e, current real-world snowfall datasets are limited in scale and lack dedicated evaluation metrics designed specifically for snowfall degradation, thus hinderin g the effective integration of real snowy images into model training to reduce d omain gaps. To address these challenges, we first introduce RealSnow10K, a large -scale, high-quality dataset consisting of over 10,000 annotated real-world snow y images. In addition, we curate a preference dataset comprising 36,000 expert-r anked image pairs, enabling the adaptation of multimodal large language models (MLLMs) to better perceive snowy image quality through our innovative Multi-Model Preference Optimization (MMPO). Finally, we propose the SnowMaster, which emplo ys MMPO-enhanced MLLM to perform accurate snowy image evaluation and pseudo-labe l filtering for semi-supervised training. Experiments demonstrate that SnowMaste r delivers superior desnowing performance under real-world conditions.
****

From Faces to Voices: Learning Hierarchical Representations for High-quality Vid eo-to-Speech

Ji-Hoon Kim, Jeongsoo Choi, Jaehun Kim, Chaeyoung Jung, Joon Son Chung; Proceedi ngs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15874-15884

The objective of this study is to generate high-quality speech from silent talki ng face videos, a task also known as video-to-speech synthesis. A significant ch allenge in video-to-speech synthesis lies in the substantial modality gap betwee n silent video and multi-faceted speech. In this paper, we propose a novel vide o-to-speech system that effectively bridges this modality gap, significantly enh ancing the quality of synthesized speech. This is achieved by learning of hierar chical representations from video to speech. Specifically, we gradually transfor m silent video into acoustic feature spaces through three sequential stages -- c ontent, timbre, and prosody modeling. In each stage, we align visual factors -- lip movements, face identity, and facial expressions -- with corresponding acous tic counterparts to ensure the seamless transformation. Additionally, to generat e realistic and coherent speech from the visual representations, we employ a flo w matching model that estimates direct trajectories from a simple prior distribu tion to the target speech distribution. Extensive experiments demonstrate that o ur method achieves exceptional generation quality comparable to real utterances, outperforming existing methods by a significant margin.
****

DFM: Differentiable Feature Matching for Anomaly Detection

Sheng Wu, Yimi Wang, Xudong Liu, Yuguang Yang, Runqi Wang, Guodong Guo, David Do

ermann, Baochang Zhang; Proceedings of the Computer Vision and Pattern Recogniti on Conference (CVPR), 2025, pp. 15224-15233
Feature matching methods for unsupervised anomaly detection have demonstrated im pressive performance. Existing methods primarily rely on self-supervised trainin g and handcrafted matching schemes for task adaptation. However, they can only a chieve an inferior feature representation for anomaly detection because the feat ure extraction and matching modules are separately trained. To address these iss ues, we propose a Differentiable Feature Matching (DFM) framework for joint opt imization of the feature extractor and the matching head. DFM transforms nearest -neighbor matching into a pooling-based module and embeds it within a Feature Ma tching Network (FMN). This design enables end-to-end feature extraction and feat ure matching module training, thus providing better feature representation for a nomaly detection tasks. DFM is generic and can be incorporated into existing fea ture-matching methods. We implement DFM with various backbones and conduct exten sive experiments across various tasks and datasets, demonstrating its effectiven ess. Notably, we achieve state-of-the-art results in the continual anomaly detec tion task with instance-AUROC improvement of up to 3.9% and pixel-AP improvement of up to 5.5%.
****

FlashGS: Efficient 3D Gaussian Splatting for Large-scale and High-resolution Ren dering
Guofeng Feng, Siyan Chen, Rong Fu, Zimu Liao, Yi Wang, Tao Liu, Boni Hu, Linning Xu, Zhilin Pei, Hengjie Li, Xiuhong Li, Ninghui Sun, Xingcheng Zhang, Bo Dai; P roceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 202 5, pp. 26652-26662
Recent advances in 3D Gaussian Splatting (3DGS) have demonstrated significant po tential over traditional rendering techniques, attracting widespread attention f rom both industry and academia. However, real-time rendering with 3DGS remains a challenging problem, particularly in large-scale, high-resolution scenes due to the presence of numerous anisotropic Gaussian representations, and it has not b een extensively explored. To address this challenge, we introduce FlashGS, an op en-source CUDA library with Python bindings, featuring comprehensive algorithm d esign and optimizations, including redundancy elimination, adaptive scheduling, and efficient pipelining. First, we eliminate substantial redundant computations through precise Gaussian intersection tests, leveraging the intrinsic mechanism of the 3DGS rasterizer. During task partitioning, we propose an adaptive schedu ling strategy that accounts for variations in Gaussian size and shape. Additiona lly, we design a multi-stage pipelining strategy for color computation in the re ndering process, further accelerating performance. We conduct an extensive evalu ation of FlashGS across a diverse range of synthetic and real-world 3D scenes, e ncompassing scene sizes of up to 2.7 km^2 cityscape and resolutions of up to ove r 4K. Our approach improves 3DGS rendering performance by an order of magnitude, achieving an average speedup of 7.2x, and rendering at a minimum of 125.9 FPS, setting a new state-of-the-art in real-time 3DGS rendering. https://github.com/I nternLandMark/FlashGS.
****

PointSR: Self-Regularized Point Supervision for Drone-View Object Detection
Weizhuo Li, Yue Xi, Wenjing Jia, Zehao Zhang, Fei Li, Xiangzeng Liu, Qiguang Mia o; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11707-11716
Point-Supervised Object Detection (PSOD) in a discriminative style has recently gained significant attention for its impressive detection performance and cost-e ffectiveness. However, accurately predicting high-quality pseudo-box labels for drone-view images, which often feature densely packed small objects, remains a c hallenge. This difficulty arises primarily from the limitation of rigid sampling strategies, which hinder the pseudo-box optimization process. To address this, we propose PointSR, an effective and robust point-supervised object detection fr amework with self-regularized sampling that integrates temporal and informative constraints throughout the pseudo-box generation process. Specifically, the fram ework comprises three key components: Temporal-Ensembling Encoder (TE Encoder),

Coarse Pseudo-box Prediction, and Pseudo-box Refinement. The TE Encoder builds a n anchor prototype library by aggregating temporal information for dynamic ancho r adjustment. In Coarse Pseudo-box Prediction, anchors are refined using the pro totype library, and a set of informative samples is collected for subsequent ref inement. During Pseudo-box Refinement, these informative negative samples are us ed to suppress low-confidence candidate positive samples, thereby improving the quality of the pseudo-boxes. Experimental results on benchmark datasets demonstr ate that PointSR significantly outperforms state-of-the-art methods, achieving u p to 2.6%~\mathbf 7.2% higher AP_ 50 using only point supervision. Additional ly, it exhibits strong robustness to perturbation in human-labeled points.
****

Exploring Timeline Control for Facial Motion Generation
Yifeng Ma, Jinwei Qi, Chaonan Ji, Peng Zhang, Bang Zhang, Zhidong Deng, Liefeng Bo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) , 2025, pp. 1940-1950
This paper introduces a new control signal for facial motion generation: timelin e control. Compared to audio and text signals, timelines provide more fine-grain ed control, such as generating specific facial motions with precise timing. User s can specify a multi-track timeline of facial actions arranged in temporal inte rvals, allowing precise control over the timing of each action. To model the tim eline control capability, We first annotate the time intervals of facial actions in natural facial motion sequences at a frame-level granularity. This process i s facilitated by Toeplitz Inverse Covariance-based Clustering to minimize human labor. Based on the annotations, we propose a diffusion-based generation model c apable of generating facial motions that are natural and accurately aligned with input timelines. Our method supports text-guided motion generation by using Cha tGPT to convert text into timelines. Experimental results show that our method c an annotate facial action intervals with satisfactory accuracy, and produces nat ural facial motions accurately aligned with timelines.
****

v-CLR: View-Consistent Learning for Open-World Instance Segmentation
Chang-Bin Zhang, Jinhong Ni, Yujie Zhong, Kai Han; Proceedings of the Computer V ision and Pattern Recognition Conference (CVPR), 2025, pp. 20307-20317
In this paper, we address the challenging problem of open-world instance segment ation. Existing works have shown that vanilla visual networks are biased toward learning appearance information, e.g. texture, to recognize objects. This implic it bias causes the model to fail in detecting novel objects with unseen textures in the open-world setting. To address this challenge, we propose a learning fra mework, called view-Consistent LeaRning (v-CLR), which aims to enforce the model to learn appearance-invariant representations for robust instance segmentation. In v-CLR, we first introduce additional views for each image, where the texture undergoes significant alterations while preserving the image's underlying struc ture. We then encourage the model to learn the appearance-invariant representati on by enforcing the consistency between object features across different views, for which we obtain class-agnostic object proposals using off-the-shelf unsuperv ised models that possess strong object-awareness. These proposals enable cross-v iew object feature matching, greatly reducing the appearance dependency while en hancing the object-awareness. We thoroughly evaluate our method on public benchm arks under both cross-class and cross-dataset settings, achieving state-of-the-a rt performance. Project page: https://visual-ai.github.io/vclr
****

Chat2SVG: Vector Graphics Generation with Large Language Models and Image Diffus ion Models
Ronghuan Wu, Wanchao Su, Jing Liao; Proceedings of the Computer Vision and Patte rn Recognition Conference (CVPR), 2025, pp. 23690-23700
Scalable Vector Graphics (SVG) has become the de facto standard for vector graph ics in digital design, offering resolution independence and precise control over individual elements. Despite their advantages, creating high-quality SVG conten t remains challenging, as it demands technical expertise with professional editi ng software and a considerable time investment to craft complex shapes. Recent t

ext-to-SVG generation methods aim to make vector graphics creation more accessib le, but they still encounter limitations in shape regularity, generalization abi lity, and expressiveness. To address these challenges, we introduce Chat2SVG, a hybrid framework that combines the strengths of Large Language Models (LLMs) and image diffusion models for text-to-SVG generation. Our approach first uses an L LM to generate semantically meaningful SVG templates from basic geometric primit ives. Guided by image diffusion models, a dual-stage optimization pipeline refin es paths in latent space and adjusts point coordinates to enhance geometric comp lexity. Extensive experiments show that Chat2SVG outperforms existing methods in visual fidelity, path regularity, and semantic alignment. Additionally, our sys tem enables intuitive editing through natural language instructions, making prof essional vector graphics creation accessible to all users. Our code is available at https://chat2svg.github.io/.

****

GAF: Gaussian Avatar Reconstruction from Monocular Videos via Multi-view Diffusi on

Jiapeng Tang, Davide Davoli, Tobias Kirschstein, Liam Schoneveld, Matthias Nießn er; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) , 2025, pp. 5546-5558

We propose a novel approach for reconstructing animatable 3D Gaussian avatars fr om monocular videos captured by commodity devices like smartphones. Photorealist ic 3D head avatar reconstruction from such recordings is challenging due to limi ted observations, which leaves unobserved regions under-constrained and can lead to artifacts in novel views. To address this problem, we introduce a multi-view head diffusion model, leveraging its priors to fill in missing regions and ensu re view consistency in Gaussian splatting renderings. To enable precise viewpoin t control, we use normal maps rendered from FLAME-based head reconstruction, whi ch provides pixel-aligned inductive biases. We also condition the diffusion mode l on VAE features extracted from the input image to preserve facial identity and appearance details. For Gaussian avatar reconstruction, we distill multi-view d iffusion priors by using iteratively denoised images as pseudo-ground truths, ef fectively mitigating over-saturation issues. To further improve photorealism, w e apply latent upsampling priors to refine the denoised latent before decoding i t into an image. We evaluate our method on the NeRSemble dataset, showing that G AF outperforms previous state-of-the-art methods in novel view synthesis. Furthe rmore, we demonstrate higher-fidelity avatar reconstructions from monocular vide os captured on commodity devices.

****

Reloc3r: Large-Scale Training of Relative Camera Pose Regression for Generalizab le, Fast, and Accurate Visual Localization

Siyan Dong, Shuzhe Wang, Shaohui Liu, Lulu Cai, Qingnan Fan, Juho Kannala, Yanch ao Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16739-16752

Visual localization aims to determine the camera pose of a query image relative to a database of posed images. In recent years, deep neural networks that direct ly regress camera poses have gained popularity due to their fast inference capab ilities. However, existing methods struggle to either generalize well to new sce nes or provide accurate camera pose estimates. To address these issues, we prese nt Reloc3r, a simple yet effective visual localization framework. It consists of an elegantly designed relative pose regression network, and a minimalist motion averaging module for absolute pose estimation. Trained on approximately eight m illion posed image pairs, Reloc3r achieves surprisingly good performance and gen eralization ability. We conduct extensive experiments on six public datasets, co nsistently demonstrating the effectiveness and efficiency of the proposed method . It provides high-quality camera pose estimates in real time and generalizes to novel scenes. Code: https://github.com/ffrivera0/reloc3r.

****

AI-Face: A Million-Scale Demographically Annotated AI-Generated Face Dataset and Fairness Benchmark

Li Lin, Santosh Santosh, Mingyang Wu, Xin Wang, Shu Hu; Proceedings of the Compu

ter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3503-3515
AI-generated faces have enriched human life, such as entertainment, education, and art. However, they also pose misuse risks. Therefore, detecting AI-generated faces becomes crucial, yet current detectors show biased performance across different demographic groups. Mitigating biases can be done by designing algorithmic fairness methods, which usually require demographically annotated face datasets for model training. However, no existing dataset encompasses both demographic attributes and diverse generative methods simultaneously, which hinders the development of fair detectors for AI-generated faces. In this work, we introduce the AI-Face dataset, the first million-scale demographically annotated AI-generated face image dataset, including real faces, faces from deepfake videos, and faces generated by Generative Adversarial Networks and Diffusion Models. Based on this dataset, we conduct the first comprehensive fairness benchmark to assess various AI face detectors and provide valuable insights and findings to promote the future fair design of AI face detectors. Our AI-Face dataset and benchmark code are publicly available at https://github.com/Purdue-M2/AI-Face-FairnessBench.
****

Inference-Scale Complexity in ANN-SNN Conversion for High-Performance and Low-Power Applications
Tong Bu, Maohua Li, Zhaofei Yu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24387-24397
Spiking Neural Networks (SNNs) have emerged as a promising substitute for Artificial Neural Networks (ANNs) due to their advantages of fast inference and low power consumption. However, the lack of efficient training algorithms has hindered their widespread adoption. Even efficient ANN-SNN conversion methods necessitate quantized training of ANNs to enhance the effectiveness of the conversion, incurring additional training costs. To address these challenges, we propose an efficient ANN-SNN conversion framework with only inference scale complexity. The conversion framework includes a local threshold balancing algorithm, which enables efficient calculation of the optimal thresholds and fine-grained adjustment of the threshold value by channel-wise scaling. We also introduce an effective delayed evaluation strategy to mitigate the influence of the spike propagation delays. We demonstrate the scalability of our framework in typical computer vision tasks: image classification, semantic segmentation, object detection, and video classification. Our algorithm outperforms existing methods, highlighting its practical applicability and efficiency. Moreover, we have evaluated the energy consumption of the converted SNNs, demonstrating their superior low-power advantage compared to conventional ANNs. This approach simplifies the deployment of SNNs by leveraging open-source pre-trained ANN models, enabling fast, low-power inference with negligible performance reduction.
****

Janus: Decoupling Visual Encoding for Unified Multimodal Understanding and Generation
Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, Ping Luo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12966-12977
We introduce Janus, an autoregressive framework that unifies multimodal understanding and generation. Prior research often relies on a single visual encoder for both tasks, such as Chameleon. However, due to the differing levels of information granularity required by multimodal understanding and generation, this approach can lead to suboptimal performance, particularly in multimodal understanding. To address this issue, we decouple visual encoding into separate pathways, while still leveraging a single, unified transformer architecture for processing. The decoupling not only alleviates the conflict between the visual encoder's roles in understanding and generation, but also enhances the framework's flexibility. For instance, both the multimodal understanding and generation components can independently select their most suitable encoding methods. Experiments show that Janus surpasses previous unified model and matches or exceeds the performance of task-specific models. The simplicity, high flexibility, and effectiveness of Janus make it a strong candidate for next-generation unified multimodal models.

****
MVDoppler-Pose: Multi-Modal Multi-View mmWave Sensing for Long-Distance Self-Occluded Human Walking Pose Estimation
Jaeho Choi, Soheil Hor, Shubo Yang, Amin Arbabian; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27750-27759
One of the main challenges in reliable camera-based 3D pose estimation for walking subjects is to deal with self-occlusions, especially in the case of using low-resolution cameras or at longer distance scenarios. In recent years, millimeter-wave (mmWave) radar has emerged as a promising alternative, offering inherent resilience to the effect of occlusions and distance variations. However, mmWave-based human walking pose estimation (HWPE) is still in the nascent development stages, primarily due to its unique set of practical challenges including the quality of the observed radar signal dependent on the subject's motion direction. This paper introduces the first comprehensive study comparing mmWave radar to camera systems for HWPE, highlighting its utility for distance-agnostic and occlusion-resilient pose estimation. Building upon mmWave's unique advantages, we address its intrinsic directionality issue through a new approach--the synergetic integration of multi-modal, multi-view mmWave signals, achieving robust HWPE against variations both in distance and walking direction. Extensive experiments on a newly curated dataset not only demonstrate the superior potential of mmWave technology over traditional camera-based HWPE systems, but also validate the effectiveness of our approach in overcoming the core limitations of mmWave HWPE.
****
TopNet: Transformer-Efficient Occupancy Prediction Network for Octree-Structured Point Cloud Geometry Compression
Xinjie Wang, Yifan Zhang, Ting Liu, Xinpu Liu, Ke Xu, Jianwei Wan, Yulan Guo, Hanyun Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27305-27314
Efficient Point Cloud Geometry Compression (PCGC) with a lower bits per point (BPP) and higher peak signal-to-noise ratio (PSNR) is essential for the transportation of large-scale 3D data. Although octree-based entropy models can reduce BPP without introducing geometry distortion, existing CNN-based models struggle with limited receptive fields to capture long-range dependencies, while Transformer-built architectures always neglect fine-grained details due to their reliance on global self-attention. In this paper, we propose a Transformer-efficient occupancy prediction Network, termed TopNet, to overcome these challenges by developing several novel components: Locally-enhanced Context Encoding (LeCE) for enhancing the translation-invariance of the octree nodes, Adaptive-Length Sliding Window Attention (AL-SWA) for capturing both global and local dependencies while adaptively adjusting attention weights based on the input window length, Spatial-Gated-enhanced Channel Mixer (SG-CM) for efficient feature aggregation from ancestors and siblings, and Latent-guided Node Occupancy Predictor (LNOP) for improving prediction accuracy of spatially adjacent octree nodes. Comprehensive experiments across both indoor and outdoor point cloud datasets demonstrate that our TopNet achieves state-of-the-art performance with fewer parameters, further advancing the reduction-efficiency boundaries of PCGC. The code is available at https://github.com/xinjiewang1995/TopNet.
****
MagicArticulate: Make Your 3D Models Articulation-Ready
Chaoyue Song, Jianfeng Zhang, Xiu Li, Fan Yang, Yiwen Chen, Zhongcong Xu, Jun Hao Liew, Xiaoyang Guo, Fayao Liu, Jiashi Feng, Guosheng Lin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15998-16007
With the explosive growth of 3D content creation, there is an increasing demand for automatically converting static 3D models into articulation-ready versions that support realistic animation. Traditional approaches rely heavily on manual annotation, which is both time-consuming and labor-intensive. Moreover, the lack of large-scale benchmarks has hindered the development of learning-based solutions. In this work, we present MagicArticulate, an effective framework that automatically transforms static 3D models into articulation-ready assets. Our key contributions are threefold. First, we introduce Articulation-XL, a large-scale benc

hmark containing over 33k 3D models with high-quality articulation annotations, carefully curated from Objaverse-XL. Second, we propose a novel skeleton generation method that formulates the task as a sequence modeling problem, leveraging an auto-regressive transformer to naturally handle varying numbers of bones or joints within skeletons and their inherent dependencies across different 3D models. Third, we predict skinning weights using a functional diffusion process that incorporates volumetric geodesic distance priors between vertices and joints. Extensive experiments demonstrate that MagicArticulate significantly outperforms existing methods across diverse object categories, achieving high-quality articulation that enables realistic animation. Project page: https://chaoyuesong.github.io/MagicArticulate.

****

Gain from Neighbors: Boosting Model Robustness in the Wild via Adversarial Perturbations Toward Neighboring Classes

Zhou Yang, Mingtao Feng, Tao Huang, Fangfang Wu, Weisheng Dong, Xin Li, Guangming Shi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25497-25507

Recent approaches, such as data augmentation, adversarial training, and transfer learning, have shown potential in addressing the issue of performance degradation caused by distributional shifts. However, they typically demand careful design in terms of data or models and lack awareness of the impact of distributional shifts. In this paper, we observe that classification errors arising from distribution shifts tend to cluster near the true values, suggesting that misclassifications commonly occur in semantically similar, neighboring categories. Furthermore, robust advanced vision foundation models maintain larger inter-class distances while preserving semantic consistency, making them less vulnerable to such shifts. Building on these findings, we propose a new method called GFN (Gain From Neighbors), which uses gradient priors from neighboring classes to perturb input images and incorporates an inter-class distance-weighted loss to improve class separation. This approach encourages the model to learn more resilient features from data prone to errors, enhancing its robustness against shifts in diverse settings. In extensive experiments across various model architectures and benchmark datasets, GFN consistently demonstrated superior performance. For instance, compared to the current state-of-the-art TAPADL method, our approach achieved a higher corruption robustness of 41.4% on ImageNet-C (+2.3%), without requiring additional parameters and using only minimal data.

****

Enhancing Video-LLM Reasoning via Agent-of-Thoughts Distillation

Yudi Shi, Shangzhe Di, Qirui Chen, Weidi Xie; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8523-8533

This paper tackles the problem of video question answering (VideoQA), a task that often requires multi-step reasoning and a profound understanding of spatial-temporal dynamics. While large video-language models perform well on benchmarks, they often lack explainability and spatial-temporal grounding. In this paper, we propose **A**gent-**o**f-**T**houghts **D**istillation (**AoTD**), a method that enhances models by incorporating automatically generated Chain-of-Thoughts (CoTs) into the instruction-tuning process. Specifically, we leverage an agent-based system to decompose complex questions into sub-tasks, and address them with specialized vision models, the intermediate results are then treated as reasoning chains. We also introduce a verification mechanism using a large language model (LLM) to ensure the reliability of generated CoTs. Extensive experiments demonstrate that AoTD improves the performance on multiple-choice and open-ended benchmarks.

****

De^2Gaze: Deformable and Decoupled Representation Learning for 3D Gaze Estimation

Yunfeng Xiao, Xiaowei Bai, Baojun Chen, Hao Su, Hao He, Liang Xie, Erwei Yin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3091-3100

3D Gaze estimation is a challenging task due to two main issues. First, existing

methods focus on analyzing dense features (e.g., large pixel regions), which are sensitive to local noise (e.g., light spots, blurs) and result in increased computational complexity. Second, an eyeball model can correspond multiple gaze directions, and the entangled representation between gazes and models increases the learning difficulty. To address these issues, we propose De\textsuperscript 2Gaze , a lightweight and accurate model-aware 3D gaze estimation method. In De\textsuperscript 2 Gaze, we introduce two key innovations for deformable and decoupled representation learning. Specifically, first, we propose a deformable sparse attention mechanism that can adapt sparse sampling points to attention areas to avoid local noise influences. Second, we propose a spatial decoupling network with a dual-branch decoding architecture to disentangle invariant (e.g., eyeball radius, position) and variable (e.g., gaze, pupil, iris) features from the latent space. Compared to existing methods, De\textsuperscript 2 Gaze requires fewer sparse features, and achieves faster convergence speed, lower computational complexity, and higher accuracy in 3D gaze estimation.Qualitative and quantitative experiments demonstrate that De\textsuperscript 2 Gaze achieves state-of-the-art accuracy and high-quality semantic segmentation for 3D gaze estimation on the TEyeD dataset.

****

ReCapture: Generative Video Camera Controls for User-Provided Videos using Masked Video Fine-Tuning

David Junhao Zhang, Roni Paiss, Shiran Zada, Nikhil Karnad, David E. Jacobs, Yael Pritch, Inbar Mosseri, Mike Zheng Shou, Neal Wadhwa, Nataniel Ruiz; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2050-2062

Recently, breakthroughs in video modeling have allowed for controllable camera trajectories in generated videos. However, these methods cannot be directly applied to user-provided videos that are not generated by a video model. In this paper, we present ReCapture, a method for generating new videos with novel camera trajectories from a single user-provided video. Our method allows us to re-generate the reference video, with all its existing scene motion, from vastly different angles and with cinematic camera motion. Notably, using our method we can also plausibly hallucinate parts of the scene that were not observable in the reference video. Our method works by (1) generating a noisy anchor video with a new camera trajectory using multiview diffusion models or depth-based point cloud rendering and then (2) regenerating the anchor video into a clean and temporally consistent reangled video using our proposed masked video fine-tuning technique.

****

M^3-VOS: Multi-Phase, Multi-Transition, and Multi-Scenery Video Object Segmentation

Zixuan Chen, Jiaxin Li, Junxuan Liang, Liming Tan, Yejie Guo, Cewu Lu, Yong-Lu Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29193-29202

Intelligent robots need to interact with diverse objects across various environments. The appearance and state of objects frequently undergo complex transformations depending on the object properties, e.g., phase transitions. However, in the vision community, segmenting dynamic objects with phase transitions is overlooked. In light of this, we introduce the concept of phase in segmentation, which categorizes real-world objects based on their visual characteristics and potential morphological and appearance changes. Then, we present a new benchmark, Multi-Phase, Multi-Transition, and Multi-Scenery Video Object Segmentation (M^3-VOS), to verify the ability of models to understand object phases, which consists of 471 high-resolution videos spanning over 10 distinct everyday scenarios. It provides dense instance mask annotations that capture both object phases and their transitions. We evaluate state-of-the-art methods on M^3-VOS, yielding several key insights. Notably, current appearance-based approaches show significant room for improvement when handling objects with phase transitions. The inherent changes in disorder suggest that the predictive performance of the forward entropy-increasing process can be improved through a reverse entropy-reducing process. These findings lead us to propose ReVOS, a new plug-and-play model that improves its

performance by reversal refinement. Our data and code will be publicly availabl
e at https://zixuan-chen.github.io/M-cube-VOS.github.io/.
****

Self-Expansion of Pre-trained Models with Mixture of Adapters for Continual Lear
ning
Huiyi Wang, Haodong Lu, Lina Yao, Dong Gong; Proceedings of the Computer Vision
and Pattern Recognition Conference (CVPR), 2025, pp. 10087-10098
Continual learning (CL) aims to continually accumulate knowledge from a non-stat
ionary data stream without catastrophic forgetting of learned knowledge, requiri
ng a balance between stability and adaptability. Relying on the generalizable re
presentation in pre-trained models (PTMs), PTM-based CL methods perform effectiv
e continual adaptation on downstream tasks by adding learnable adapters or promp
ts upon the frozen PTMs. However, many existing PTM-based CL methods use restric
ted adaptation on a fixed set of these modules to avoid forgetting, suffering fr
om limited CL ability. Periodically adding task-specific modules results in line
ar model growth rate and impaired knowledge reuse. We propose Self-Expansion of
pre-trained models with Modularized Adaptation (SEMA), a novel approach to enhan
ce the control of stability-plasticity balance in PTM-based CL. SEMA automatical
ly decides to reuse or add adapter modules on demand in CL, depending on whether
significant distribution shift that cannot be handled is detected at different
representation levels. We design modular adapter consisting of a functional adap
ter and a representation descriptor. The representation descriptors are trained
as a distribution shift indicator and used to trigger self-expansion signals. Fo
r better composing the adapters, an expandable weighting router is learned joint
ly for mixture of adapter outputs. SEMA enables better knowledge reuse and sub-l
inear expansion rate. Extensive experiments demonstrate the effectiveness of the
proposed self-expansion method, achieving state-of-the-art performance compared
to PTM-based CL methods without memory rehearsal. Code is available at https://
github.com/huiyiwang01/SEMA-CL.
****

Dual Prompting Image Restoration with Diffusion Transformers
Dehong Kong, Fan Li, Zhixin Wang, Jiaqi Xu, Renjing Pei, Wenbo Li, WenQi Ren; Pr
oceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025
, pp. 12809-12819
Recent state-of-the-art image restoration methods mostly adopt latent diffusion
models with U-Net backbones, yet still facing challenges in achieving high-quali
ty restoration due to their limited capabilities. Diffusion transformers (DiTs),
like SD3, are emerging as a promising alternative because of their better quali
ty with scalability. However, previous conditional control methods for U-Net-bas
ed diffusion models, such as ControlNet, are not well-suited for DiTs. In this p
aper, we introduce DPIR (Dual Recent state-of-the-art image restoration methods
mostly adopt latent diffusion models with U-Net backbones, yet still facing chal
lenges in achieving high-quality restoration due to their limited capabilities.
Diffusion transformers (DiTs), like SD3, are emerging as a promising alternative
because of their better quality with scalability. In this paper, we introduce D
PIR (Dual Prompting Image Restoration), a novel image restoration method that ef
fectivly extracts conditional information of low-quality images from multiple pe
rspectives. Specifically, DPIR consits of two branches: a low-quality image cond
itioning branch and a dual prompting control branch. The first branch utilizes a
lightweight module to incorporate image priors into the DiT with high efficienc
y. More importantly, we believe that in image restoration, textual description a
lone cannot fully capture its rich visual characteristics. Therefore, a dual pro
mpting module is designed to provide DiT with additional visual cues, capturing
both global context and local appearance. The extracted global-local visual prom
pts as extra conditional control, alongside textual prompts to form dual prompts
, greatly enhance the quality of the restoration. Extensive experimental results
demonstrate that DPIR delivers superior image restoration performance.
****

Brain-Inspired Spiking Neural Networks for Energy-Efficient Object Detection
Ziqi Li, Tao Gao, Yisheng An, Ting Chen, Jing Zhang, Yuanbo Wen, Mengkun Liu, Qi

anxi Zhang; Proceedings of the Computer Vision and Pattern Recognition Conferenc e (CVPR), 2025, pp. 3552-3562
Brain-inspired spiking neural networks (SNNs) have the capability of energy-effi cient processing of temporal information. However, leveraging the rich dynamic c haracteristics of SNNs and prior works in artificial neural networks (ANNs) to c onstruct an effective object detection model for visual tasks remains an open qu estion for further exploration. To develop a directly-trained , low energy consu mption and high-performance multi-scale SNN model, we propose a novel interpreta ble object detection framework Multi-scale Spiking Detector (MSD). Initially, we propose a spiking convolutional neuron as a core component of the Optic Nerve N ucleus Block (ONNB), designed to significantly enhance the deep feature extracti on capabilities of SNNs. ONNB enables direct training with improved energy effic iency, demonstrating superior performance compared to state-of-the-art ANN-to-SN N conversion and SNN techniques. In addition, we propose a Multi-scale Spiking D etection Framework to emulate the biological response and comprehension of stimu li from different objects. Wherein, spiking multi-scale fusion and the spiking d etector are employed to integrate features across different depths and to detect response outcomes, respectively. Our method outperforms state-of-the-art ANN de tectors, with only 7.8 M parameters and 6.43 mJ energy consumption. MSD obtains the mean average precision (mAP) of 62.0% and 66.3% on COCO and Gen1 datasets, r espectively.
****

Medusa: A Multi-Scale High-order Contrastive Dual-Diffusion Approach for Multi-V iew Clustering
Liang Chen, Zhe Xue, Yawen Li, Meiyu Liang, Yan Wang, Anton van den Hengel, Yuan kai Qi; Proceedings of the Computer Vision and Pattern Recognition Conference (C VPR), 2025, pp. 10295-10304
Deep multi-view clustering methods utilize information from multiple views to ac hieve enhanced clustering results and have gained increasing popularity in recen t years. Most existing methods typically focus on either inter-view or intra-vie w relationships, aiming to align information across views or analyze structural patterns within individual views. However, they often incorporate inter-view com plementary information in a simplistic manner, while overlooking the complex, hi gh-order relationships within multi-view data and the interactions among samples , resulting in an incomplete utilization of the rich information available. Inst ead, we propose a multi-scale approach that exploits all of the available inform ation. We first introduce a dual graph diffusion module guided by a consensus gr aph. This module leverages inter-view information to enhance the representation of both nodes and edges within each view. Secondly, we propose a novel contrasti ve loss function based on hypergraphs to more effectively model and leverage com plex intra-view data relationships. Finally, we propose to adaptively learn fusi on weights at the sample level, which enables a more flexible and dynamic aggreg ation of multi-view information. Extensive experiments on eight datasets show fa vorable performance of the proposed method compared to state-of-the-art approach es, demonstrating its effectiveness across diverse scenarios.
****

MambaOut: Do We Really Need Mamba for Vision?
Weihao Yu, Xinchao Wang; Proceedings of the Computer Vision and Pattern Recognit ion Conference (CVPR), 2025, pp. 4484-4496
Mamba, an architecture with RNN-like token mixer of state space model (SSM), was recently introduced to address the quadratic complexity of the attention mechan ism and subsequently applied to vision tasks. Nevertheless, the performance of M amba for vision is often underwhelming when compared with convolutional and atte ntion-based models. In this paper, we delve into the essence of Mamba, and conce ptually conclude that Mamba is ideally suited for tasks with long-sequence and a utoregressive characteristics. For vision tasks, as image classification on Imag eNet does not align with either characteristic, we hypothesize that Mamba is not necessary for this task; Detection and segmentation tasks on COCO or ADE20K are also not autoregressive, yet they adhere to the long-sequence characteristic, s o we believe it is still worthwhile to explore Mamba's potential for these tasks

. To empirically verify our hypotheses, we construct a series of models named MambaOut through stacking Mamba blocks while removing their core token mixer, SSM. Experimental results strongly support our hypotheses. Specifically, our MambaOut model surpasses all visual Mamba models on ImageNet image classification, indicating that Mamba is indeed unnecessary for this task. As for detection and segmentation, MambaOut cannot match the performance of state-of-the-art visual Mamba models, demonstrating the potential of Mamba for long-sequence visual tasks.
****

Everything to the Synthetic: Diffusion-driven Test-time Adaptation via Synthetic-Domain Alignment

Jiayi Guo, Junhao Zhao, Chaoqun Du, Yulin Wang, Chunjiang Ge, Zanlin Ni, Shiji Song, Humphrey Shi, Gao Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30503-30513

Test-time adaptation (TTA) aims to improve the performance of source-domain pre-trained models on previously unseen, shifted target domains. Traditional TTA methods primarily adapt model weights based on target data streams, making model performance sensitive to the amount and order of target data. The recently proposed diffusion-driven TTA methods mitigate this by adapting model inputs instead of weights, where an unconditional diffusion model, trained on the source domain, transforms target-domain data into a synthetic domain that is expected to approximate the source domain. However, in this paper, we reveal that although the synthetic data in diffusion-driven TTA seems indistinguishable from the source data, it is unaligned with, or even markedly different from the latter for deep networks. To address this issue, we propose a Synthetic-Domain Alignment (SDA) framework. Our key insight is to fine-tune the source model with synthetic data to ensure better alignment. Specifically, we first employ a conditional diffusion model to generate labeled samples, creating a synthetic dataset. Subsequently, we use the aforementioned unconditional diffusion model to add noise to and denoise each sample before fine-tuning. This Mix of Diffusion (MoD) process mitigates the potential domain misalignment between the conditional and unconditional models. Extensive experiments across classifiers, segmenters, and multimodal large language models (MLLMs, e.g., LLaVA) demonstrate that SDA achieves superior domain alignment and consistently outperforms existing diffusion-driven TTA methods. Our code is available at https://github.com/SHI-Labs/Diffusion-Driven-Test-Time-Adaptation-via-Synthetic-Domain-Alignment.
****

Multi-Granularity Class Prototype Topology Distillation for Class-Incremental Source-Free Unsupervised Domain Adaptation

Peihua Deng, Jiehua Zhang, Xichun Sheng, Chenggang Yan, Yaoqi Sun, Ying Fu, Liang Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30566-30576

This paper explores the Class-Incremental Source-Free Unsupervised Domain Adaptation (CI-SFUDA) problem, where the unlabeled target data come incrementally without access to labeled source instances. This problem poses two challenges, the interference of similar source-class knowledge in target-class representation learning and the shocks of new target knowledge to old ones. To address them, we propose the Multi-Granularity Class Prototype Topology Distillation (GROTO) algorithm, which effectively transfers the source knowledge to the class-incremental target domain. Concretely, we design the multi-granularity class prototype self-organization module and the prototype topology distillation module. First, we mine the positive classes by modeling accumulation distributions. Next, we introduce multi-granularity class prototypes to generate reliable pseudo-labels, and exploit them to promote the positive-class target feature self-organization. Second, the positive-class prototypes are leveraged to construct the topological structures of source and target feature spaces. Then, we perform the topology distillation to continually mitigate the shocks of new target knowledge to old ones. Extensive experiments demonstrate that our proposed method achieves state-of-the-art performance on three public datasets.
****

DepthCues: Evaluating Monocular Depth Perception in Large Vision Models

Duolikun Danier, Mehmet Aygün, Changjian Li, Hakan Bilen, Oisin Mac Aodha; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20049-20059

Large-scale pre-trained vision models are becoming increasingly prevalent, offering expressive and generalizable visual representations that benefit various downstream tasks. Recent studies on the emergent properties of these models have revealed their high-level geometric understanding, in particular in the context of depth perception. However, it remains unclear how depth perception arises in these models without explicit depth supervision provided during pre-training. To investigate this, we examine whether the monocular depth cues, similar to those used by the human visual system, emerge in these models. We introduce a new benchmark, DepthCues, designed to evaluate depth cue understanding, and present findings across 20 diverse and representative pre-trained vision models. Our analysis shows that human-like depth cues emerge in more recent larger models. We also explore enhancing depth perception in large vision models by fine-tuning on DepthCues, and find that even without dense depth supervision, this improves depth estimation. To support further research, our benchmark and evaluation code will be made publicly available for studying depth perception in vision models.

****

A Polarization-Aided Transformer for Image Deblurring via Motion Vector Decomposition

Duosheng Chen, Shihao Zhou, Jinshan Pan, Jinglei Shi, Lishen Qu, Jufeng Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28061-28070

Effectively leveraging motion information is crucial for the image deblurring task. Existing methods typically build deep-learning models to restore a clean image by estimating blur patterns over the entire movement. This suggests that the blur caused by rotational motion components is processed together with the translational one. Exploring the movement without separation leads to limited performance for complex motion deblurring, especially rotational motion. In this paper, we propose Motion Decomposition Transformer (MDT), a transformer-based architecture augmented with polarized modules for deblurring via motion vector decomposition. MDT consists of a Motion Decomposition Module (MDM) for extracting hybrid rotation and translation features, and a Radial Stripe Attention Solver (RSAS) for sharp image reconstruction with enhanced rotational information. Specifically, the MDM uses a deformable Cartesian convolutional branch to capture translational motion, complemented by a polar-system branch to capture rotational motion. The RSAS employs radial stripe windows and angular relative positional encoding in the polar system to enhance rotational information. This design preserves translational details while keeping computational costs lower than dual-coordinate design. Experimental results on 6 image deblurring datasets show that MDT outperforms state-of-the-art methods, particularly in handling blur caused by complex motions with significant rotational components.

****

SpecTRe-GS: Modeling Highly Specular Surfaces with Reflected Nearby Objects by Tracing Rays in 3D Gaussian Splatting

Jiajun Tang, Fan Fei, Zhihao Li, Xiao Tang, Shiyong Liu, Youyu Chen, Binxiao Huang, Zhenyu Chen, Xiaofei Wu, Boxin Shi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16133-16142

3D Gaussian Splatting (3DGS), a recently emerged multi-view 3D reconstruction technique, has shown significant advantages in real-time rendering and explicit editing. However, 3DGS encounters challenges in the accurate modeling of both high-frequency view-dependent appearances and global illumination effects, including inter-reflection. This paper introduces SpecTRe-GS, which addresses these challenges and models highly Specular surfaces that reflect nearby objects through Tracing Rays in 3D Gaussian Splatting. SpecTRe-GS separately models reflections from highly specular and rough surfaces to leverage the distinctions between their reflective properties and integrates an efficient ray tracer within the 3DGS framework for querying secondary rays, thus achieving fast and accurate rendering. Also, it incorporates normal prior guidance and joint geometry optimization at

various stages of the training process to enhance geometry reconstruction for un
distorted reflections. Experiments on both synthetic and real-world scenes demon
strate the superiority of SpecTRe-GS compared to existing 3DGS-based methods in
capturing highly specular inter-reflections and also showcase its editing applic
ations.
****

Seurat: From Moving Points to Depth
Seokju Cho, Jiahui Huang, Seungryong Kim, Joon-Young Lee; Proceedings of the Com
puter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7211-7221
Accurate depth estimation from monocular videos remains challenging due to ambig
uities inherent in single-view geometry, as crucial depth cues like stereopsis a
re absent. However, humans often perceive relative depth intuitively by observin
g variations in the size and spacing of objects as they move. Inspired by this,
we propose a novel method that infers relative depth by examining the spatial re
lationships and temporal evolution of a set of tracked 2D trajectories. Specific
ally, we use off-the-shelf point tracking models to capture 2D trajectories. The
n, our approach employs spatial and temporal transformers to process these traje
ctories and directly infer depth changes over time. Evaluated on the TAPVid-3D b
enchmark, our method demonstrates robust zero-shot performance, generalizing eff
ectively from synthetic to real-world datasets. Results indicate that our approa
ch achieves temporally smooth, high-accuracy depth predictions across diverse do
mains.
****

AuraFusion360: Augmented Unseen Region Alignment for Reference-based 360deg Unbo
unded Scene Inpainting
Chung-Ho Wu, Yang-Jung Chen, Ying-Huan Chen, Jie-Ying Lee, Bo-Hsu Ke, Chun-Wei T
uan Mu, Yi-Chuan Huang, Chin-Yang Lin, Min-Hung Chen, Yen-Yu Lin, Yu-Lun Liu; Pr
oceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025
, pp. 16366-16376
Three-dimensional scene inpainting is crucial for applications from virtual real
ity to architectural visualization, yet existing methods struggle with view cons
istency and geometric accuracy in 360deg unbounded scenes. We present AuraFusion
360, a novel reference-based method that enables high-quality object removal and
 hole filling in 3D scenes represented by Gaussian Splatting. Our approach intro
duces (1) depth-aware unseen mask generation for accurate occlusion identificati
on, (2) Adaptive Guided Depth Diffusion, a zero-shot method for accurate initial
 point placement without requiring additional training, and (3) SDEdit-based det
ail enhancement for multi-view coherence. We also introduce 360-USID, the first
comprehensive dataset for 360deg unbounded scene inpainting with ground truth. E
xtensive experiments demonstrate that AuraFusion360 significantly outperforms ex
isting methods, achieving superior perceptual quality while maintaining geometri
c accuracy across dramatic viewpoint changes.
****

Language-Guided Image Tokenization for Generation
Kaiwen Zha, Lijun Yu, Alireza Fathi, David A. Ross, Cordelia Schmid, Dina Katabi
, Xiuye Gu; Proceedings of the Computer Vision and Pattern Recognition Conferenc
e (CVPR), 2025, pp. 15713-15722
Image tokenization, the process of transforming raw image pixels into a compact
low-dimensional latent representation, has proven crucial for scalable and effic
ient image generation. However, mainstream image tokenization methods generally
have limited compression rates, making high-resolution image generation computat
ionally expensive. To address this challenge, we propose to leverage language fo
r efficient image tokenization, and we call our method Text-Conditioned Image To
kenization (TexTok). TexTok is a simple yet effective tokenization framework tha
t leverages language to provide a compact, high-level semantic representation. B
y conditioning the tokenization process on descriptive text captions, TexTok sim
plifies semantic learning, allowing more learning capacity and token space to be
 allocated to capture fine-grained visual details, leading to enhanced reconstru
ction quality and higher compression rates. Compared to the conventional tokeniz
er without text conditioning, TexTok achieves average reconstruction FID improve

ments of 29.2% and 48.1% on ImageNet-256 and -512 benchmarks respectively, acros s varying numbers of tokens. These tokenization improvements consistently transl ate to 16.3% and 34.3% average improvements in generation FID. By simply replaci ng the tokenizer in Diffusion Transformer (DiT) with TexTok, our system can achi eve a 93.5x inference speedup while still outperforming the original DiT using o nly 32 tokens on ImageNet-512. TexTok with a vanilla DiT generator achieves stat e-of-the-art FID scores of 1.46 and 1.62 on ImageNet-256 and -512 respectively. Furthermore, we demonstrate TexTok's superiority on the text-to-image generation task, effectively utilizing the off-the-shelf text captions in tokenization.
****

Img-Diff: Contrastive Data Synthesis for Multimodal Large Language Models
Qirui Jiao, Daoyuan Chen, Yilun Huang, Bolin Ding, Yaliang Li, Ying Shen; Procee dings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp . 9296-9307
High-performance Multimodal Large Language Models (MLLMs) rely heavily on data q uality. This study introduces a novel data synthesis method, leveraging insights from contrastive learning and image difference captioning to enhance fine-grain ed image recognition in MLLMs. By analyzing object differences in detailed regio ns between similar images, we challenge the model to identify both matching and distinct components. Specifically, our method initially create pairs of similar images that highlight object variations. After that, we introduce a Difference A rea Generator for object differences identifying, followed by a Difference Capti ons Generator for differences describing. The outcome is a high-quality dataset of "object replacement" samples, named Img-Diff, which can be expanded as needed due to its automation. We use the generated dataset to finetune state-of-the-ar t (SOTA) MLLMs such as InternVL2, yielding comprehensive improvements across num erous image difference and Visual Question Answering tasks. For instance, the tr ained models notably surpass the SOTA models GPT-4V and Gemini on the MMVP bench mark. Additionally, we conduct thorough evaluations to confirm the dataset's div ersity, quality, and robustness, presenting several insights on the synthesis of such a contrastive dataset. We release our codes and dataset to encourage furth er research on multimodal data synthesis and MLLMs' fundamental capabilities for image understanding.
****

CocoER: Aligning Multi-Level Feature by Competition and Coordination for Emotio n Recognition
Xuli Shen, Hua Cai, Weilin Shen, Qing Xu, Dingding Yu, Weifeng Ge, Xiangyang Xue ; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29591-29600
With the explosion of human-machine interaction, emotion recognition has reigni ted attention. Previous works focus on improving visual feature fusion and reas oning from multiple image levels. Although it is non-trivial to deduce a person 's emotion by integrating multi-level feature (head, body and context), the emo tion recognition results of each level is usually different from one another, wh ich creates inconsistency in the prevailing feature alignment method and decreas e recognition performance. In this work, we propose a multi-level image feature refinement method for emotion recognition (CocoER) to mitigate the impact caused by conflicting results from multi-level recognition. First, we leverage cross-level attention to improve visual feature consistency between hierarchically cr opped head, body and context windows. Then, vocabulary informed alignment is in corporated into the recognition framework to produce pseudo label and guide hier archical visual feature refinement. To effectively fuse multi-level feature, we elaborate on a competition process of eliminating irrelevant image level predic tions and a coordination process to enhance the feature across all levels. Exte nsive experiments are executed on two popular datasets, and our method achieves state-of-the-art performance with multi-level interpretation results.
****

Hyperbolic Uncertainty-Aware Few-Shot Incremental Point Cloud Segmentation
Tanuj Sur, Samrat Mukherjee, Kaizer Rahaman, Subhasis Chaudhuri, Muhammad Haris Khan, Biplab Banerjee; Proceedings of the Computer Vision and Pattern Recognitio

n Conference (CVPR), 2025, pp. 11810-11821
3D point cloud segmentation is essential across a range of applications; however, conventional methods often struggle in evolving environments, particularly when tasked with identifying novel categories under limited supervision. Few-Shot Learning (FSL) and Class Incremental Learning (CIL) have been adapted previously to address these challenges in isolation, yet the combined paradigm of Few-Shot Class Incremental Learning (FSCIL) remains largely unexplored for point cloud segmentation. To address this gap, we introduce Hyperbolic Ideal Prototypes Optimization (HIPO), a novel framework that harnesses hyperbolic embeddings for FSCIL in 3D point clouds. HIPO employs the Poincare Hyperbolic Sphere as its embedding space, integrating Ideal Prototypes enriched by CLIP-derived class semantics, to capture the hierarchical structure of 3D data. By enforcing orthogonality among prototypes and maximizing representational margins, HIPO constructs a resilient embedding space that mitigates forgetting and enables the seamless integration of new classes, thereby effectively countering overfitting. Extensive evaluations on S3DIS, ScanNetv2, and cross-dataset scenarios demonstrate HIPO's strong performance, significantly surpassing existing approaches in both in-domain and cross-dataset FSCIL tasks for 3D point cloud segmentation.
****

Enhancing Creative Generation on Stable Diffusion-based Models
Jiyeon Han, Dahee Kwon, Gayoung Lee, Junho Kim, Jaesik Choi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28609-28618
Recent text-to-image generative models, particularly Stable Diffusion and its distilled variants, have achieved impressive fidelity and strong text-image alignment. However, their creative generation capacity remains limited, as simply adding the term "creative" to prompts often fails to yield genuinely creative results. In this paper, we introduce C3 (Creative Concept Catalyst), a training-free approach designed to enhance creativity in Stable Diffusion-based models. C3 selectively amplifies features during the denoising process to foster more creative outputs. We offer practical guidelines for choosing amplification factors based on two main aspects of creativity. C3 allows user-friendly creativity control in image generation and is the first study to enhance creativity in diffusion models without extensive computational costs. We demonstrate its effectiveness across various Stable Diffusion-based models. Source codes will be publicly available.
****

The Devil is in the Prompts: Retrieval-Augmented Prompt Optimization for Text-to-Video Generation
Bingjie Gao, Xinyu Gao, Xiaoxue Wu, Yujie Zhou, Yu Qiao, Li Niu, Xinyuan Chen, Yaohui Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3173-3183
The evolution of Text-to-video (T2V) generative models, trained on large-scale datasets, has been marked by significant progress. However, the sensitivity of T2V generative models to input prompts highlights the critical role of prompt design in influencing generative outcomes. Prior research has predominantly relied on Large Language Models (LLMs) to align user-provided prompts with the distribution of training prompts, albeit without tailored guidance encompassing prompt vocabulary and sentence structure nuances. To this end, we introduce RAPO, a novel Retrieval-Augmented Prompt Optimization framework. In order to address potential inaccuracies and ambiguous details generated by LLM-generated prompts. RAPO refines the naive prompts through dual optimization branches, selecting the superior prompt for T2V generation. The first branch augments user prompts with diverse modifiers extracted from a learned relational graph, refining them to align with the format of training prompts via a fine-tuned LLM. Conversely, the second branch rewrites the naive prompt using a pre-trained LLM following a well-defined instruction set. Extensive experiments demonstrate that RAPO can effectively enhance both the static and dynamic dimensions of generated videos, demonstrating the significance of prompt optimization for user-provided prompts. Project website: \href https://whynothaha.github.io/Prompt_optimizer/RAPO.html GitHub .
****

Denoising Functional Maps: Diffusion Models for Shape Correspondence
Aleksei Zhuravlev, Zorah Lähner, Vladislav Golyanik; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26899-26909
Estimating correspondences between pairs of deformable shapes remains a challenging problem. Despite substantial progress, existing methods lack broad generalization capabilities and require category-specific training data. To address these limitations, we propose a fundamentally new approach to shape correspondence based on denoising diffusion models. In our method, a diffusion model learns to directly predict the functional map, a low-dimensional representation of a point-wise map between shapes. We use a large dataset of synthetic human meshes for training and employ two steps to reduce the number of functional maps that need to be learned. First, the maps refer to a template rather than shape pairs. Second, the functional map is defined in a basis of eigenvectors of the Laplacian, which is not unique due to sign ambiguity. Therefore, we introduce an unsupervised approach to select a specific basis by correcting the signs of eigenvectors based on surface features. Our model achieves competitive performance on standard human datasets, meshes with anisotropic connectivity, non-isometric humanoid shapes, as well as animals compared to existing descriptor-based and large-scale shape deformation methods.
****

ProReflow: Progressive Reflow with Decomposed Velocity
Lei Ke, Haohang Xu, Xuefei Ning, Yu Li, Jiajun Li, Haoling Li, Yuxuan Lin, Dongsheng Jiang, Yujiu Yang, Linfeng Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28029-28038
Diffusion models have achieved significant progress in both image and video generation while still suffering from huge computation costs. As an effective solution, rectified flow aims to rectify the diffusion process of diffusion models into a straight line for few-step and even one-step generation. However, in this paper, we suggest that the original training pipeline of reflow is not optimal and introduce two techniques to improve it. Firstly, we introduce progressive reflow, which progressively reflows the diffusion models in local timesteps until the whole diffusion progresses, reducing the difficulty of flow matching. Second, we introduce aligned v-prediction, which highlights the importance of direction matching in flow matching over magnitude matching. Experimental results on SDv1.5 and SDXL demonstrate the effectiveness of our method, for example, conducting on SDv1.5 achieves an FID of 10.70 on MSCOCO2014 validation set with only 4 sampling steps, close to our teacher model (32 DDIM steps, FID = 10.05). Our codes will be released at Github.
****

DnLUT: Ultra-Efficient Color Image Denoising via Channel-Aware Lookup Tables
Sidi Yang, Binxiao Huang, Yulun Zhang, Dahai Yu, Yujiu Yang, Ngai Wong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7582-7591
While deep neural networks have revolutionized image denoising capabilities, their deployment on edge devices remains challenging due to substantial computational and memory requirements. To this end, we present DnLUT, an ultra-efficient lookup table-based framework that achieves high-quality color image denoising with minimal resource consumption. Our key innovation lies in two complementary components: a Pairwise Channel Mixer (PCM) that effectively captures inter-channel correlations and spatial dependencies in parallel, and a novel L-shaped convolution design that maximizes receptive field coverage while minimizing storage overhead. By converting these components into optimized lookup tables post-training, DnLUT achieves remarkable efficiency - requiring only 500KB storage and 0.1% energy consumption compared to its CNN contestant DnCNN, while delivering 20x faster inference. Extensive experiments demonstrate that DnLUT outperforms all existing LUT-based methods by over 1dB in PSNR, establishing a new state-of-the-art in resource-efficient color image denoising.
****

Devil is in the Detail: Towards Injecting Fine Details of Image Prompt in Image Generation via Conflict-free Guidance and Stratified Attention

Kyungmin Jo, Jooyeol Yun, Jaegul Choo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23595-23603

While large-scale text-to-image diffusion models enable the generation of high-quality, diverse images from text prompts, these prompts struggle to capture intricate details, such as textures, preventing the user intent from being reflected. This limitation has led to efforts to generate images conditioned on user-provided images, referred to as image prompts. Recent work modifies the self-attention mechanism to impose image conditions in generated images by replacing or concatenating the keys and values from the image prompt. This enables the self-attention layer to work like a cross-attention layer, generally used to incorporate text prompts.In this paper, we identify two common issues in existing methods of modifying self-attention that hinder diffusion models from reflecting the image prompt. By addressing these issues, we propose a novel method that generates images that properly reflect the details of image prompts. First, existing approaches often neglect the importance of image prompts in classifier-free guidance, which directs the model towards the intended conditions and away from those undesirable. Specifically, current methods use image prompts as both desired and undesired conditions, causing conflicting signals. To resolve this, we propose conflict-free guidance by using image prompts only as desired conditions, ensuring that the generated image faithfully reflects the image prompt.In addition, we observe that the two most common self-attention modifications involve a trade-off between the realism of the generated image and alignment with the image prompt, achieved by selectively using keys and values from both images. Specifically, selecting more keys and values from the image prompt improves alignment, while selecting more from the generated image enhances realism. To balance both, we propose an alternative self-attention modification method, Stratified Attention, which jointly uses keys and values from both images rather than selecting between them. Through extensive experiments across three distinct image generation tasks, we demonstrate that the proposed method outperforms existing image-prompting models in faithfully reflecting the image prompt.

****

D^3-Human: Dynamic Disentangled Digital Human from Monocular Video

Honghu Chen, Bo Peng, Yunfan Tao, Juyong Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10836-10846

We introduce \text D ^3\text -Human , a method for reconstructing Dynamic Disentangled Digital Human geometry from monocular videos. Past monocular video human reconstruction primarily focuses on reconstructing undecoupled clothed human bodies or only reconstructing clothing, making it difficult to apply directly in applications such as animation production. The challenge in reconstructing decoupled clothing and body lies in the occlusion caused by clothing over the body. To this end, the details of the visible area and the plausibility of the invisible area must be ensured during the reconstruction process. Our proposed method combines explicit and implicit representations to model the decoupled clothed human body, leveraging the robustness of explicit representations and the flexibility of implicit representations. Specifically, we reconstruct the visible region as SDF and propose a novel human manifold signed distance field (hmSDF) to segment the visible clothing and visible body, and then merge the visible and invisible body. Extensive experimental results demonstrate that, compared with existing reconstruction schemes, \text D ^3\text -Human can achieve high-quality decoupled reconstruction of the human body wearing different clothing, and can be directly applied to clothing transfer and animation production.

****

BiM-VFI: Bidirectional Motion Field-Guided Frame Interpolation for Video with Non-uniform Motions

Wonyong Seo, Jihyong Oh, Munchurl Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7244-7253

Existing Video Frame interpolation (VFI) models tend to suffer from time-to-location ambiguity when trained with video of non-uniform motions, such as accelerating, decelerating, and changing directions, which often yield blurred interpolated frames.In this paper, we propose (i) a novel motion description map, Bidirect

ional Motion field (BiM), to effectively describe non-uniform motions; (ii) a BiM-guided Flow Net (BiMFN) with Content-Aware Upsampling Network (CAUN) for precise optical flow estimation; and (iii) Knowledge Distillation for VFI-centric Flow supervision (KDVCF) to supervise the motion estimation of VFI model with VFI-centric teacher flows.The proposed VFI is called a Bidirectional Motion field-guided VFI (BiM-VFI) model.Extensive experiments show that our BiM-VFI model significantly surpasses the recent state-of-the-art VFI methods by 26% and 45% improvements in LPIPS and STLPIPS respectively, yielding interpolated frames with much fewer blurs at arbitrary time instances.
****

Curriculum Coarse-to-Fine Selection for High-IPC Dataset Distillation
Yanda Chen, Gongwei Chen, Miao Zhang, Weili Guan, Liqiang Nie; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20437-20446
Dataset distillation (DD) excels in synthesizing a small number of images per class (IPC) but struggles to maintain its effectiveness in high-IPC settings. Recent works on dataset distillation demonstrate that combining distilled and real data can mitigate the effectiveness decay. However, our analysis of the combination paradigm reveals that the current one-shot and independent selection mechanism induces an incompatibility issue between distilled and real images. To address this issue, we introduce a novel curriculum coarse-to-fine selection (CCFS) method for efficient high-IPC dataset distillation. CCFS employs a curriculum selection framework for real data selection, where we leverage a coarse-to-fine strategy to select appropriate real data based on the current synthetic dataset in each curriculum. Extensive experiments validate CCFS, surpassing the state-of-the-art by +6.6% on CIFAR-10, +5.8% on CIFAR-100, and +3.4% on Tiny-ImageNet under high-IPC settings. Notably, CCFS achieves 60.2% test accuracy on ResNet-18 with a 20% compression ratio of Tiny-ImageNet, closely matching full-dataset training with only 0.3% degradation. Code: https://github.com/CYDaaa30/CCFS.
****

BADGR: Bundle Adjustment Diffusion Conditioned by Gradients for Wide-Baseline Floor Plan Reconstruction
Yuguang Li, Ivaylo Boyadzhiev, Zixuan Liu, Linda Shapiro, Alex Colburn; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16785-16795
Reconstructing precise camera poses and floor plan layouts from wide-baseline RGB panoramas is a difficult and unsolved problem. We introduce BADGR, a novel diffusion model that jointly performs reconstruction and bundle adjustment (BA) to refine poses and layouts from a coarse state, using 1D floor boundary predictions from dozens of sparsely captured images. Unlike guided diffusion models, BADGR is conditioned on dense per-column outputs from a single-step Levenberg Marquardt (LM) optimizer and is trained to predict camera and wall positions, while minimizing reprojection errors for view consistency. The objective of layout generation from denoising diffusion process complements BA optimization by providing additional learned layout-structural constraints on top of the co-visible features across images. These constraints help BADGR make plausible guesses about spatial relationships, which constrain the pose graph, such as wall adjacency and collinearity, while also learning to mitigate errors from dense boundary observations using global context. BADGR trains exclusively on 2D floor plans, simplifying data acquisition, enabling robust augmentation, and supporting a variety of input densities. Our experiments validate our method, which significantly outperforms the state-of-the-art pose and floor plan layout reconstruction with different input densities.
****

Three Cars Approaching within 100m! Enhancing Distant Geometry by Tri-Axis Voxel Scanning for Camera-based Semantic Scene Completion
Jongseong Bae, Junwoo Ha, Ha Young Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11939-11948
Camera-based Semantic Scene Completion (SSC) is gaining attentions in the 3D perception field. However, properties such as perspective and occlusion lead to the

underestimation of the geometry in distant regions, posing a critical issue for safety-focused autonomous driving systems. To tackle this, we propose ScanSSC, a novel camera-based SSC model composed of a Scan Module and Scan Loss, both designed to enhance distant scenes by leveraging context from near-viewpoint scenes. The Scan Module uses axis-wise masked attention, where each axis employing a near-to-far cascade masking that enables distant voxels to capture relationships with preceding voxels. In addition, the Scan Loss computes the cross-entropy along each axis between cumulative logits and corresponding class distributions in a near-to-far direction, thereby propagating rich context-aware signals to distant voxels. Leveraging the synergy between these components, ScanSSC achieves state-of-the-art performance, with IoUs of 44.54 and 48.29, and mIoUs of 17.40 and 20.14 on the SemanticKITTI and SSCBench-KITTI-360 benchmarks.

****

MetaShadow: Object-Centered Shadow Detection, Removal, and Synthesis

Tianyu Wang, Jianming Zhang, Haitian Zheng, Zhihong Ding, Scott Cohen, Zhe Lin, Wei Xiong, Chi-Wing Fu, Luis Figueroa, Soo Ye Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28252-28262

Shadows are often underconsidered or even ignored in image editing applications, limiting the realism of the edited results. In this paper, we introduce MetaShadow, a three-in-one versatile framework that enables detection, removal, and controllable synthesis of shadows in natural images in an object-centered fashion. MetaShadow combines the strengths of two cooperative components: Shadow Analyzer, for object-centered shadow detection and removal, and Shadow Synthesizer, for reference-based controllable shadow synthesis. Notably, we optimize the learning of the intermediate features from Shadow Analyzer to guide Shadow Synthesizer to generate more realistic shadows that blend seamlessly with the scene. Extensive evaluations on multiple shadow benchmark datasets show significant improvements of MetaShadow over the existing state-of-the-art methods on object-centered shadow detection, removal, and synthesis. MetaShadow excels in supporting imageediting tasks such as object removal, relocation, and insertion, pushing the boundaries of object-centered image editing.

****

TANGO: Training-free Embodied AI Agents for Open-world Tasks

Filippo Ziliotto, Tommaso Campari, Luciano Serafini, Lamberto Ballan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24603-24613

Large Language Models (LLMs) have demonstrated excellent capabilities in composing various modules together to create programs that can perform complex reasoning tasks on images. In this paper, we propose TANGO, an approach that extends the program composition via LLMs already observed for images, aiming to integrate those capabilities into embodied agents capable of observing and acting in the world. Specifically, by employing a simple PointGoal Navigation model combined with a memory-based exploration policy as a foundational primitive for guiding an agent through the world, we show how a single model can address diverse tasks without additional training. We task an LLM with composing the provided primitives to solve a specific task, using only a few in-context examples in the prompt. We evaluate our approach on three key Embodied AI tasks: Open-Set ObjectGoal Navigation, Multi-Modal Lifelong Navigation, and Open Embodied Question Answering, achieving state-of-the-art results without any specific fine-tuning in challenging zero-shot scenarios.

****

SATA: Spatial Autocorrelation Token Analysis for Enhancing the Robustness of Vision Transformers

Nick Nikzad, Yi Liao, Yongsheng Gao, Jun Zhou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9730-9739

Over the past few years, vision transformers (ViTs) have consistently demonstrated remarkable performance across various visual recognition tasks. However, attempts to enhance their robustness have yielded limited success, mainly focusing on different training strategies, input patch augmentation, or network structural enhancements. These approaches often involve extensive training and fine-tuning

, which are time-consuming and resource-intensive. To tackle these obstacles, we introduce a novel approach named Spatial Autocorrelation Token Analysis (SATA). By harnessing spatial relationships between token features, SATA enhances both the representational capacity and robustness of ViT models. This is achieved through the analysis and grouping of tokens according to their spatial autocorrelation scores prior to their input into the Feed-Forward Network (FFN) block of the self-attention mechanism. Importantly, SATA seamlessly integrates into existing pre-trained ViT baselines without requiring retraining or additional fine-tuning, while concurrently improving efficiency by reducing the computational load of the FFN units. Experimental results show that the baseline ViTs enhanced with SATA not only achieve a new state-of-the-art top-1 accuracy on ImageNet-1K image classification (94.9%) but also establish new state-of-the-art performance across multiple robustness benchmarks, including ImageNet-A (top-1=63.6%), ImageNet-R (top-1=79.2%), and ImageNet-C (mCE=13.6%), all without requiring additional training or fine-tuning of baseline models. Availability: https://github.com/nick-nikzad/SATA

****

DViN: Dynamic Visual Routing Network for Weakly Supervised Referring Expression Comprehension

Xiaofu Chen, Yaxin Luo, Gen Luo, Jiayi Ji, Henghui Ding, Yiyi Zhou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14347-14357

In this paper, we focus on weakly supervised referring expression comprehension (REC), and identify that the lack of fine-grained visual capability greatly limits the upper performance bound of existing methods. To address this issue, we propose a novel framework for weakly supervised REC, namely Dynamic Visual routing Network (DViN), which overcomes the visual shortcomings from the perspective of feature combination and alignment. In particular, DViN is equipped with a novel sparse routing mechanism to efficiently combine features of multiple visual encoders in a dynamic manner, thus improving the visual descriptive power. Besides, we further propose an innovative weakly supervised objective, namely Routing-based Feature Alignment (RFA), which facilitates the visual understanding of routed features through the intra-modal and inter-modal alignment. To validate DViN, we conduct extensive experiments on four REC benchmark datasets. Experiments demonstrate that DViN achieves state-of-the-art results on four benchmarks while maintaining competitive inference efficiency. Besides, the strong generalization ability of DViN is also validated on weakly supervised referring expression segmentation. Source codes are anonymously released at: https://anonymous.4open.science/r/DViN-7736.

****

Nested Diffusion Models Using Hierarchical Latent Priors

Xiao Zhang, Ruoxi Jiang, Rebecca Willett, Michael Maire; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2502-2512

We introduce nested diffusion models, an efficient and powerful hierarchical generative framework that substantially enhances the generation quality of diffusion models, particularly for images of complex scenes. Our approach employs a series of diffusion models to progressively generate latent variables at different semantic levels. Each model in this series is conditioned on the output of the preceding higher-level models, culminating in image generation. Hierarchical latent variables guide the generation process along predefined semantic pathways, allowing our approach to capture intricate structural details. To construct these latent variables, we leverage a pre-trained visual encoder, which learns strong semantic visual representations, and modulate its capacity via dimensionality reduction and noise injection. Across multiple datasets, our system demonstrates significant enhancements in image quality for both unconditional and class/text conditional generation. Moreover, our unconditional generation system substantially outperforms the baseline conditional system. These advancements incur minimal computational overhead as the more abstract levels of our hierarchy work with lower-dimensional representations.

****

A Theory of Learning Unified Model via Knowledge Integration from Label Space Varying Domains
Dexuan Zhang, Thomas Westfechtel, Tatsuya Harada; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10142-10152
Existing domain adaptation systems can hardly be applied to real-world problems with new classes presenting at deployment time, especially regarding source-free scenarios where multiple source domains do not share the label space despite being given a few labeled target data. To address this, we consider a challenging problem: multi-source semi-supervised open-set domain adaptation and propose a learning theory via joint error, effectively tackling strong domain shift. To generalize the algorithm into source-free cases, we introdcue a computationally efficient and architecture-flexible attention-based feature generation module. Extensive experiments on various data sets demonstrate the significant improvement of our proposed algorithm over baselines.
****

HiLoTs: High-Low Temporal Sensitive Representation Learning for Semi-Supervised LiDAR Segmentation in Autonomous Driving
R.D. Lin, Pengcheng Weng, Yinqiao Wang, Han Ding, Jinsong Han, Fei Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1429-1438
LiDAR point cloud semantic segmentation plays a crucial role in autonomous driving. In recent years, semi-supervised methods have gained popularity due to their significant reduction in annotation labor and time costs. Current semi-supervised methods typically focus on point cloud spatial distribution or consider short-term temporal representations, e.g., only two adjacent frames, often overlooking the rich long-term temporal properties inherent in autonomous driving scenarios. In driving experience, we observe that nearby objects, such as roads and vehicles, remain stable while driving, whereas distant objects exhibit greater variability in category and shape. This natural phenomenon is also captured by LiDAR, which reflects lower temporal sensitivity for nearby objects and higher sensitivity for distant ones. To leverage these characteristics, we propose HiLoTs, which learns high-temporal sensitivity and low-temporal sensitivity representations from continuous LiDAR frames. These representations are further enhanced and fused using a cross-attention mechanism. Additionally, we employ a teacher-student framework to align the representations learned by the labeled and unlabeled branches, effectively utilizing the large amounts of unlabeled data. Experimental results on the SemanticKITTI and nuScenes datasets demonstrate that our proposed HiLoTs outperforms state-of-the-art semi-supervised methods, and achieves performance close to LiDAR+Camera multimodal approaches.
****

Spiking Transformer with Spatial-Temporal Attention
Donghyun Lee, Yuhang Li, Youngeun Kim, Shiting Xiao, Priyadarshini Panda; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13948-13958
Spike-based Transformer presents a compelling and energy-efficient alternative to traditional Artificial Neural Network (ANN)-based Transformers, achieving impressive results through sparse binary computations. However, existing spike-based transformers predominantly focus on spatial attention while neglecting crucial temporal dependencies inherent in spike-based processing, leading to suboptimal feature representation and limited performance. To address this limitation, we propose Spiking Transformer with Spatial-Temporal Attention (STAtten), a simple and straightforward architecture that efficiently integrates both spatial and temporal information in the self-attention mechanism. STAtten introduces a block-wise computation strategy that processes information in spatial-temporal chunks, enabling comprehensive feature capture while maintaining the same computational complexity as previous spatial-only approaches. Our method can be seamlessly integrated into existing spike-based transformers without architectural overhaul. Extensive experiments demonstrate that STAtten significantly improves the performance of existing spike-based transformers across both static and neuromorphic datasets, including CIFAR10/100, ImageNet, CIFAR10-DVS, and N-Caltech101.

****
Perceptual Video Compression with Neural Wrapping
Muhammad Umar Karim Khan, Aaron Chadha, Mohammad Ashraful Anam, Yiannis Andreopoulos; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17743-17754
Standard video codecs are rate-distortion optimization machines, where distortion is typically quantified using PSNR versus the source. However, it is now widely accepted that increasing PSNR does not necessarily translate to better visual quality. In this paper, a better balance between perception and fidelity is targeted, in order to provide for significant rate savings over state-of-the-art standards-based video codecs. Specifically, pre- and postprocessing neural networks are proposed that enhance the coding efficiency of standard video codecs when benchmarked with an array of well-established perceptual quality scores. These "neural wrapper" elements are end-to-end trained with a neural codec module serving as a differentiable proxy for standard video codecs. The codec proxy is jointly optimized with the pre- and post components, via a novel two-phase pretraining strategy and end-to-end iterative refinement with stop-gradient. This allows the neural pre- and postprocessor to learn to embed, remove and recover information in a codec-aware manner, thus improving its rate-quality performance. A single neural-wrapper model is thereby established and used for the entire rate-quality curve without needing any downscaling or upscaling. The trained model is tested with the AV1 and VVC standard codecs via an array of well-established objective quality scores (SSIM, MS-SSIM, VMAF, AVQT), as well as mean opinion scores (MOS) derived from ITU-T P.910 subjective testing. Experimental results show that the proposed approach improves all quality scores, with -18.5% average Bjontegaard Delta-rate (BD-rate) saving over all objective scores and MOS improvement over both standard codecs. This illustrates the significant potential of neural wrapper components over standards-based video coding.
****
ViKIENet: Towards Efficient 3D Object Detection with Virtual Key Instance Enhanced Network
Zhuochen Yu, Bijie Qiu, Andy W. H. Khong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11844-11853
The sparsity of point clouds and inadequacy of semantic information pose challenges to current LiDAR-only 3D object detection methods. Recent methods alleviate these challenges by converting RGB images into virtual points via depth completion to be fused with LiDAR points. Although these methods have shown outstanding results, they often introduce significant computation overhead due to the high density of virtual points and noise due to inaccurate depth completion. Besides, they do not thoroughly leverage semantic information from images. In this work, we propose the virtual key instance enhanced network (ViKIENet), a highly efficient and effective multi-modal feature fusion framework that fuses the features of virtual key instances (VKIs) and LiDAR points through multiple stages. Our contributions include three main components: semantic key instance selection (SKIS), virtual-instance-focused fusion (VIFF), and virtual-instance-to-real attention (VIRA). We also propose the extended version ViKIENet-R with VIFF-R which includes rotationally equivariant features. Experiment results show that ViKIENet and ViKIENet-R achieve significant improvements in detection performance on the KITTI, JRDB, and nuScenes datasets compared to existing works. On the KITTI dataset, ViKIENet and ViKIENet-R operate at 22.7 and 15.0 FPS, respectively. As of CVPR submission (Nov. 15th, 2024), ViKIENet ranks first on the car detection and orientation estimation leaderboard, while ViKIENet-R ranks second (compared with officially published papers) on the 3D car detection leaderboard.
****
DKDM: Data-Free Knowledge Distillation for Diffusion Models with Any Architecture
Qianlong Xiang, Miao Zhang, Yuzhang Shang, Jianlong Wu, Yan Yan, Liqiang Nie; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2955-2965
Diffusion models (DMs) have demonstrated exceptional generative capabilities acr

oss various domains, including image, video, and so on. A key factor contributin g to their effectiveness is the high quantity and quality of data used during tr aining. However, mainstream DMs now consume increasingly large amounts of data. For example, training a Stable Diffusion model requires billions of image-text p airs. This enormous data requirement poses significant challenges for training l arge DMs due to high data acquisition costs and storage expenses. To alleviate t his data burden, we propose a novel scenario: using existing DMs as data sources to train new DMs with any architecture. We refer to this scenario as Data-Free Knowledge Distillation for Diffusion Models (DKDM), where the generative ability of DMs is transferred to new ones in a data-free manner. To tackle this challen ge, we make two main contributions. First, we introduce a DKDM objective that en ables the training of new DMs via distillation, without requiring access to the data. Second, we develop a dynamic iterative distillation method that efficientl y extracts time-domain knowledge from existing DMs, enabling direct retrieval of training data without the need for a prolonged generative process. To the best of our knowledge, we are the first to explore this scenario. Experimental result s demonstrate that our data-free approach not only achieves competitive generati ve performance but also, in some instances, outperforms models trained with the entire dataset.

****

SymDPO: Boosting In-Context Learning of Large Multimodal Models with Symbol Demo nstration Direct Preference Optimization

Hongrui Jia, Chaoya Jiang, Haiyang Xu, Wei Ye, Mengfan Dong, Ming Yan, Ji Zhang, Fei Huang, Shikun Zhang; Proceedings of the Computer Vision and Pattern Recogni tion Conference (CVPR), 2025, pp. 9361-9371

As language models continue to scale, Large Language Models (LLMs) have exhibite d emerging capabilities in In-Context Learning (ICL), enabling them to solve lan guage tasks by prefixing a few in-context demonstrations (ICDs) as context. Insp ired by these advancements, researchers have extended these techniques to develo p Large Multimodal Models (LMMs) with ICL capabilities. However, existing LMMs f ace a critical issue: they often fail to effectively leverage the visual context in multimodal demonstrations and instead simply follow textual patterns. This i ndicates that LMMs do not achieve effective alignment between multimodal demonst rations and model outputs. To address this problem, we propose Symbol Demonstrat ion Direct Preference Optimization (SymDPO). Specifically, SymDPO aims to break the traditional paradigm of constructing multimodal demonstrations by using rand om symbols to replace text answers within instances. This forces the model to ca refully understand the demonstration images and establish a relationship between the images and the symbols to answer questions correctly. We validate the effec tiveness of this method on multiple benchmarks, demonstrating that with SymDPO, LMMs can more effectively understand the multimodal context within examples and utilize this knowledge to answer questions better. Code is available at https:// github.com/APiaoG/SymDPO.

****

Stealthy Backdoor Attack in Self-Supervised Learning Vision Encoders for Large V ision Language Models

Zhaoyi Liu, Huan Zhang; Proceedings of the Computer Vision and Pattern Recogniti on Conference (CVPR), 2025, pp. 25060-25070

Self-supervised learning (SSL) vision encoders learn high-quality image represen tations and thus have become a vital part of developing vision modality of large vision language models (LVLMs). Due to the high cost of training such encoders, pre-trained encoders are widely shared and deployed into many LVLMs, which are security-critical or bear societal significance. Under this practical scenario, we reveal a new backdoor threat that significant visual hallucinations can be in duced into these LVLMs by merely compromising vision encoders. Because of the sh aring and reuse of these encoders, many downstream LVLMs may inherit backdoor be haviors from encoders, leading to widespread backdoors. In this work, we propose BadVision, the first method to exploit this vulnerability in SSL vision encoder s for LVLMs with novel trigger optimization and backdoor learning techniques. We evaluate BadVision on two types of SSL encoders and LVLMs across eight benchmar

ks. We show that BadVision effectively drives the LVLMs to attacker-chosen hallu cination with over 99% attack success rate, causing a 77.6% relative visual unde rstanding error while maintaining the stealthiness. SoTA backdoor detection meth ods cannot detect our attack effectively.
****

Data-free Universal Adversarial Perturbation with Pseudo-semantic Prior
Chanhui Lee, Yeonghwan Song, Jeany Son; Proceedings of the Computer Vision and P attern Recognition Conference (CVPR), 2025, pp. 13907-13916
Data-free Universal Adversarial Perturbation (UAP) is an image-agnostic adversar ial attack that deceives deep neural networks using a single perturbation genera ted solely from random noise without relying on data priors. However, traditiona l data-free UAP methods often suffer from limited transferability due to the abs ence of semantic content in random noise. To address this issue, we propose a no vel data-free universal attack method that recursively extracts pseudo-semantic priors directly from the UAPs during training to enrich the semantic content wit hin the data-free UAP framework. Our approach effectively leverages latent seman tic information within UAPs via region sampling, enabling successful input trans formations--typically ineffective in traditional data-free UAP methods due to th e lack of semantic cues--and significantly enhancing black-box transferability. Furthermore, we introduce a sample reweighting technique to mitigate potential i mbalances from random sampling and transformations, emphasizing hard examples le ss affected by the UAPs. Comprehensive experiments on ImageNet show that our met hod achieves state-of-the-art performance in average fooling rate by a substanti al margin, notably improves attack transferability across various CNN architectu res compared to existing data-free UAP methods, and even surpasses data-dependen t UAP methods. Code is available at: https://github.com/ChnanChan/PSP-UAP.
****

Debiasing Multimodal Large Language Models via Noise-Aware Preference Optimizati on
Zefeng Zhang, Hengzhu Tang, Jiawei Sheng, Zhenyu Zhang, Yiming Ren, Zhenyang Li, Dawei Yin, Duohe Ma, Tingwen Liu; Proceedings of the Computer Vision and Patter n Recognition Conference (CVPR), 2025, pp. 9423-9433
Multimodal Large Language Models (MLLMs) excel in various tasks, yet often strug gle with modality bias, tending to rely heavily on a single modality or prior kn owledge when generating responses. In this paper, we propose a debiased preferen ce optimization dataset, RLAIF-V-Bias, and introduce a Noise-Aware Preference Op timization (NAPO) algorithm. Specifically, we first construct the dataset by int roducing perturbations to reduce the informational content of certain modalities , prompting the model to overly rely on a specific modality when generating resp onses. To address the inevitable noise in automatically constructed data, we com bine the noise-robust Mean Absolute Error (MAE) with the Binary Cross-Entropy (B CE) in Direct Preference Optimization (DPO) using a negative Box-Cox transformat ion and dynamically adjust the algorithm's noise robustness based on the evaluat ed noise levels in the data.Extensive experiments validate our approach, demonst rating not only its effectiveness in mitigating modality bias but also its signi ficant role in minimizing hallucinations.
****

SAM2-LOVE: Segment Anything Model 2 in Language-aided Audio-Visual Scenes
Yuji Wang, Haoran Xu, Yong Liu, Jiaze Li, Yansong Tang; Proceedings of the Compu ter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28932-28941
Reference Audio-Visual Segmentation (Ref-AVS) aims to provide a pixel-wise scene understanding in Language-aided Audio-Visual Scenes (LAVS). This task requires the model to continuously segment objects referred to by text and audio from a v ideo. Previous dual-modality methods always fail due to the lack of a third moda lity and the existing triple-modality method struggles with spatio-temporal cons istency, leading to the target shift of different frames. In this work, we intro duce a novel framework, termed SAM2-LOVE, which integrates textual, audio, and v isual representations into a learnable token to prompt and align SAM2 for achiev ing Ref-AVS in the LAVS. Technically, our approach includes a multimodal fusion module aimed at improving multimodal understanding of SAM2, as well as token pro

pagation and accumulation strategies designed to enhance spatio-temporal consistency without forgetting historical information. We conducted extensive experiments to demonstrate that SAM2-LOVE outperforms the SOTA by 8.5% in J&F on the Ref-AVS benchmark and showcase the simplicity and effectiveness of the components. Our code will be available here.

****

GIVEPose: Gradual Intra-class Variation Elimination for RGB-based Category-Level Object Pose Estimation

Ziqin Huang, Gu Wang, Chenyangguang Zhang, Ruida Zhang, Xiu Li, Xiangyang Ji; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22055-22066

Recent advances in RGBD-based category-level object pose estimation have been limited by their reliance on precise depth information, restricting their broader applicability. In response, RGB-based methods have been developed. Among these methods, geometry-guided pose regression that originated from instance-level tasks has demonstrated strong performance. However, we argue that the NOCS map is an inadequate intermediate representation for geometry-guided pose regression method, as its many-to-one correspondence with category-level pose introduces redundant instance-specific information, resulting in suboptimal results. This paper identifies the intra-class variation problem inherent in pose regression based solely on the NOCS map and proposes the Intra-class Variation-Free Consensus (IVFC) map, a novel coordinate representation generated from the category-level consensus model. By leveraging the complementary strengths of the NOCS map and the IVFC map, we introduce GIVEPose, a framework that implements Gradual Intra-class Variation Elimination for category-level object pose estimation. Extensive evaluations on both synthetic and real-world datasets demonstrate that GIVEPose significantly outperforms existing state-of-the-art RGB-based approaches, achieving substantial improvements in category-level object pose estimation. Our code is available at https://github.com/ziqin-h/GIVEPose.

****

FRAME: Floor-aligned Representation for Avatar Motion from Egocentric Video

Andrea Boscolo Camiletto, Jian Wang, Eduardo Alvarado, Rishabh Dabral, Thabo Beeler, Marc Habermann, Christian Theobalt; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17497-17507

Egocentric motion capture with a head-mounted body-facing stereo camera is crucial for VR and AR applications but presents significant challenges such as heavy occlusions and limited annotated real-world data. Existing methods rely on synthetic pretraining and struggle to generate smooth and accurate predictions in real-world settings, particularly for lower limbs. Our work addresses these limitations by introducing a lightweight VR-based data collection setup with on-board, real-time 6D pose tracking. Using this setup, we collected the most extensive real-world dataset for ego-facing ego-mounted cameras to date in size and motion variability. Effectively integrating this multimodal input -- device pose and camera feeds -- is challenging due to the differing characteristics of each data source. To address this, we propose FRAME, a simple yet effective architecture that combines device pose and camera feeds for state-of-the-art body pose prediction through geometrically sound multimodal integration and can run at 300 FPS on modern hardware. Lastly, we showcase a novel training strategy to enhance the model's generalization capabilities. Our approach exploits the problem's geometric properties, yielding high-quality motion capture free from common artifacts in prior works. Qualitative and quantitative evaluations, along with extensive comparisons, demonstrate the effectiveness of our method. Data, code, and CAD designs will be available at https://vcai.mpi-inf.mpg.de/projects/FRAME/

****

Sketch Down the FLOPs: Towards Efficient Networks for Human Sketch

Aneeshan Sain, Subhajit Maity, Pinaki Nath Chowdhury, Shubhadeep Koley, Ayan Kumar Bhunia, Yi-Zhe Song; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28383-28393

As sketch research has collectively matured over time, its adaptation for at-mass commercialisation emerges on the immediate horizon. Despite an already mature

research endeavour for photos, there is no research on the efficient inference specifically designed for sketch data. In this paper, we first demonstrate existing state-of-the-art efficient light-weight models designed for photos do not work on sketches. We then propose two sketch-specific components which work in a plug-n-play manner on any photo efficient network to adapt them to work on sketch data. We specifically chose fine-grained sketch-based image retrieval (FG-SBIR) as a demonstrator as the most recognised sketch problem with immediate commercial value. Technically speaking, we first propose a cross-modal knowledge distillation network to transfer existing photo efficient networks to be compatible with sketch, which brings down number of FLOPs and model parameters by 97.96% percent and 84.89% respectively. We then exploit the abstract trait of sketch to introduce a RL-based canvas selector that dynamically adjusts to the abstraction level which further cuts down number of FLOPs by two thirds. The end result is an overall reduction of 99.37% of FLOPs (from 40.18G to 0.254G) when compared with a full network, while retaining the accuracy (33.03% vs 32.77%) -- finally making an efficient network for the sparse sketch data that exhibit even fewer FLOPs than the best photo counterpart.

****

Generalized Zero-Shot Classification via Semantics-Free Inter-Class Feature Generation

Libiao Chen, Dong Nie, Junjun Pan, Jing Yan, Zhenyu Tang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20286-20295

Generalized Zero-Shot Learning (GZSL) addresses the challenge of classifying unseen classes in the presence of seen classes by leveraging semantic attributes to bridge the gap for unseen classes. However, in image based disease classification, such as glioma sub-typing, distinguishing between classes using image semantic attributes can be challenging. To address this challenge, we introduce a novel GZSL method that eliminates the dependency on semantic information. Specifically, we propose that the primary of most classification in clinic is risk stratification, and classes are inherently ordered rather than purely categorical. Based on this insight, we present an inter-class feature augmentation (IFA) module, where distributions of different classes are ordered by their risk levels in a learned feature space using pre-defined joint conditional Gaussian distribution model. This ordering enables the generation of unseen class features through feature mixing of adjacent seen classes, effectively transforming the zero-shot learning problem into a supervised learning task. Our method eliminates the need for explicit semantic information, avoiding the cross-modal alignment between visual and semantic features. Moreover, the IFA module for GZSL requires no structural modifications to the existing classification models. In the experiment, both in-house and public datasets are used to evaluate our method across different tasks, including glioma subtyping, Alzheimer's disease (AD) classification and diabetic retinopathy classification. Experimental results demonstrate that our method outperforms the state-of-the-art GZSL methods with statistical significance.

****

Feat2GS: Probing Visual Foundation Models with Gaussian Splatting

Yue Chen, Xingyu Chen, Anpei Chen, Gerard Pons-Moll, Yuliang Xiu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6348-6361

Given that visual foundation models (VFMs) are trained on extensive datasets but often limited to 2D images, a natural question arises: how well do they understand the 3D world? With the differences in architecture and training protocols (i.e., objectives, proxy tasks), a unified framework to fairly and comprehensively probe their 3D awareness is urgently needed. Existing works on 3D probing suggest single-view 2.5D estimation (e.g., depth and normal) or two-view sparse 2D correspondence (e.g., matching and tracking). Unfortunately, these tasks ignore texture awareness, and require 3D data as ground-truth, which limits the scale and diversity of their evaluation set. To address these issues, we introduce Feat2GS, which readout 3D Gaussians attributes from VFM features extracted from unposed images. This allows us to probe 3D awareness for geometry and texture via novel view synthesis, without requiring 3D data. Additionally, the disentanglement o

f 3DGS parameters - geometry (x, a, S) and texture (c) - enables separate analysis of texture and geometry awareness. Under Feat2GS, we conduct extensive experiments to probe the 3D awareness of several VFMs, and investigate the ingredients that lead to a 3D aware VFM. Building on these findings, we develop several variants that achieve state-of-the-art across diverse datasets. This makes Feat2GS useful for probing VFMs, and as a simple-yet-effective baseline for novel-view synthesis. Code and data will be made available at fanegg.github.io/Feat2GS.
****

Multi-Modal Aerial-Ground Cross-View Place Recognition with Neural ODEs
Sijie Wang, Rui She, Qiyu Kang, Siqi Li, Disheng Li, Tianyu Geng, Shangshu Yu, Wee Peng Tay; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11717-11728
Place recognition (PR) aims at retrieving the query place from a database and plays a crucial role in various applications, including navigation, autonomous driving, and augmented reality. While previous multi-modal PR works have mainly focused on the same-view scenario in which ground-view descriptors are matched with a database of ground-view descriptors during inference, the multi-modal cross-view scenario, in which ground-view descriptors are matched with aerial-view descriptors in a database, remains under-explored. We propose AGPlace, a model that effectively integrates information from multi-modal ground sensors (cameras and LiDARs) to achieve accurate aerial-ground PR. AGPlace achieves effective aerial-ground cross-view PR by leveraging a manifold-based neural ordinary differential equation (ODE) framework with a multi-domain alignment loss. It outperforms existing state-of-the-art cross-view PR models on large-scale datasets. As most existing PR models are designed for ground-ground PR, we adapt these baselines into our cross-view pipeline. Experiments demonstrate that this direct adaptation performs worse than our overall model architecture AGPlace. AGPlace represents a significant advancement in multi-modal aerial-ground PR, with promising implications for real-world applications.
****

Rethinking Decoder Design: Improving Biomarker Segmentation Using Depth-to-Space Restoration and Residual Linear Attention
Saad Wazir, Daeyoung Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30861-30871
Segmenting biomarkers in medical images is crucial for various biotech applications. Despite advances, Transformer and CNN based methods often struggle with variations in staining and morphology, limiting feature extraction. In medical image segmentation, where datasets often have limited sample availability, recent state-of-the-art (SOTA) methods achieve higher accuracy by leveraging pre-trained encoders, whereas end-to-end methods tend to underperform. This is due to challenges in effectively transferring rich multiscale features from encoders to decoders, as well as limitations in decoder efficiency. To address these issues, we propose an architecture that captures multi-scale local and global contextual information and a novel decoder design, which effectively integrates features from the encoder, emphasizes important channels and regions, and reconstructs spatial dimensions to enhance segmentation accuracy. Our method, compatible with various encoders, outperforms SOTA methods, as demonstrated by experiments on four datasets and ablation studies. Specifically, our method achieves absolute performance gains of 2.76% on MoNuSeg, 3.12% on DSB, 2.87% on Electron Microscopy, and 4.03% on TNBC datasets compared to existing SOTA methods. Code: https://github.com/saadwazir/MCADS-Decoder
****

MaDCoW: Marginal Distortion Correction for Wide-Angle Photography with Arbitrary Objects
Kevin Zhang, Jia-Bin Huang, Jose Echevarria, Stephen DiVerdi, Aaron Hertzmann; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10923-10932
We introduce MaDCoW, a method for correcting marginal distortion of arbitrary objects in wide-angle photography. People often use wide-angle photography to convey natural scenes--smartphones typically default to wide-angle photography--but

depicting very wide-field-of-view scenes produces distorted object appearance, particularly marginal distortion in linear projections. With MaDCoW, a user annotates regions-of-interest to correct, along with straight lines. For each region, MaDCoW solves for a local-linear perspective projection and then jointly solves for a projection for the whole photograph that minimizes distortion. We show that our method can produce good results in cases where previous methods yield visible distortions.
****

SynTab-LLaVA: Enhancing Multimodal Table Understanding with Decoupled Synthesis
Bangbang Zhou, Zuan Gao, Zixiao Wang, Boqiang Zhang, Yuxin Wang, Zhineng Chen, Hongtao Xie; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24796-24806
Due to the limited scale of multimodal table understanding (MTU) data, model performance is constrained. A straightforward approach is to use multimodal large language models to obtain more samples, but this may cause hallucinations, generate incorrect sample pairs, and cost significantly.To address the above issues, we design a simple yet effective synthesis framework that consists of two independent steps: table image rendering and table question and answer (Q&A) pairs generation.We use table codes (HTML, LaTeX, Markdown) to synthesize images and generate Q&A pairs with large language model (LLM).This approach leverages LLM's high concurrency and low cost to boost annotation efficiency and reduce expenses. By inputting code instead of images, LLMs can directly access the content and structure of the table, reducing hallucinations in table understanding and improving the accuracy of generated Q&A pairs. Finally, we synthesize a large-scale MTU dataset, SynTab, containing 636K images and 1.8M samples costing within \200 in US dollars. We further introduce a generalist tabular multimodal model, SynTab-LLaVA. This model not only effectively extracts local textual content within the table but also enables global modeling of relationships between cells.SynTab-LLaVA achieves SOTA performance on 21 out of 24 in-domain and out-of-domain benchmarks, demonstrating the effectiveness and generalization of our method. The Code is available at \href https://github.com/bang123-box/SynTab-LLaVA SynTab-LLaVA .
****

Edit Away and My Face Will not Stay: Personal Biometric Defense against Malicious Generative Editing
Hanhui Wang, Yihua Zhang, Ruizheng Bai, Yue Zhao, Sijia Liu, Zhengzhong Tu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23806-23816
Recent advancements in diffusion models have made generative image editing more accessible than ever. While these developments allow users to generate creative edits with ease, they also raise significant ethical concerns, particularly regarding malicious edits to human portraits that threaten individuals' privacy and identity security. Existing general-purpose image protection methods primarily focus on generating adversarial perturbations to nullify edit effects. However, these approaches often exhibit instability to protect against diverse editing requests. In this work, we introduce a novel perspective to personal human portrait protection against malicious editing. Unlike traditional methods aiming to prevent edits from taking effect, our method, FaceLock, optimizes adversarial perturbations to ensure that original biometric information---such as facial features---is either destroyed or substantially altered post-editing, rendering the subject in the edited output biometrically unrecognizable. Our approach innovatively integrates facial recognition and visual perception factors into the perturbation optimization process, ensuring robust protection against a variety of editing attempts. Besides, we shed light on several critical issues with commonly used evaluation metrics in image editing and reveal cheating methods by which they can be easily manipulated, leading to deceptive assessments of protection. Through extensive experiments, we demonstrate that FaceLock significantly outperforms all baselines in defense performance against a wide range of malicious edits. Moreover, our method also exhibits strong robustness against purification techniques. Comprehensive ablation studies confirm the stability and broad applicability of our method across diverse diffusion-based editing algorithms. Our work not onl

y advances the state-of-the-art in biometric defense but also sets the foundatio
n for more secure and privacy-preserving practices in image editing.
****

Any6D: Model-free 6D Pose Estimation of Novel Objects
Taeyeop Lee, Bowen Wen, Minjun Kang, Gyuree Kang, In So Kweon, Kuk-Jin Yoon; Pro
ceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025,
 pp. 11633-11643
We introduce Any6D, a model-free framework for 6D object pose estimation that re
quires only a single RGB-D anchor image to estimate both the 6D pose and size of
 unknown objects in novel scenes. Unlike existing methods that rely on textured
3D models or multiple viewpoints, Any6D leverages a joint object alignment proce
ss to enhance 2D-3D alignment and metric scale estimation for improved pose accu
racy. Our approach integrates a render-and-compare strategy to generate and refi
ne pose hypotheses, enabling robust performance in scenarios with occlusions, no
n-overlapping views, diverse lighting conditions, and large cross-environment va
riations. We evaluate our method on five challenging datasets: REAL275, Toyota-L
ight, HO3D, YCBINEOAT, and LM-O, demonstrating its effectiveness in significantl
y outperforming state-of-the-art methods for novel object pose estimation. Proje
ct page: https://taeyeop.com/any6d
****

Improving Accuracy and Calibration via Differentiated Deep Mutual Learning
Han Liu, Peng Cui, Bingning Wang, Weipeng Chen, Yupeng Zhang, Jun Zhu, Xiaolin H
u; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR),
 2025, pp. 25812-25821
Deep Neural Networks (DNNs) have achieved remarkable success in a variety of tas
ks, particularly in terms of prediction accuracy. However, in real-world scenari
os, especially in safety-critical applications, accuracy alone is insufficient;
reliable uncertainty estimates are essential. Modern DNNs, often trained with cr
oss-entropy loss, tend to exhibit overconfidence, especially on ambiguous sample
s. Many techniques aim to improve uncertainty calibration, yet they often come a
t the cost of reduced accuracy or increased computational demands. To address th
is challenge, we propose Differentiated Deep Mutual Learning (Diff-DML), an effi
cient ensemble approach that simultaneously enhances accuracy and uncertainty ca
libration. Diff-DML draws inspiration from Deep Mutual Learning (DML) while intr
oducing two strategies to maintain prediction diversity: (1) Differentiated Trai
ning Strategy (DTS) and (2) Diversity-Preserving Learning Objective (DPLO). Our
theoretical analysis shows that Diff-DML's diversified learning framework not on
ly leverages ensemble benefits but also avoids the loss of prediction diversity
observed in traditional DML setups, which is crucial for improved calibration. E
xtensive evaluations on various benchmarks confirm the effectiveness of Diff-DML
. For instance, on the CIFAR-100 dataset, Diff-DML on ResNet34 model achieved su
bstantial improvements over the previous state-of-the-art method, MDCA, with abs
olute accuracy gains of 1.3%/3.1%, relative ECE reductions of 49.6%/43.8%, and r
elative classwise-ECE reductions of 7.7%/13.0%.
****

DrVideo: Document Retrieval Based Long Video Understanding
Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, Jianfe
i Cai; Proceedings of the Computer Vision and Pattern Recognition Conference (CV
PR), 2025, pp. 18936-18946
Most of the existing methods for video understanding primarily focus on videos o
nly lasting tens of seconds, with limited exploration of techniques for handling
 long videos. The increased number of frames in long videos poses two main chall
enges: difficulty in locating key information and performing long-range reasonin
g. Thus, we propose DrVideo, a document-retrieval-based system designed for long
 video understanding. Our key idea is to convert the long-video understanding pr
oblem into a long-document understanding task so as to effectively leverage the
power of large language models. Specifically, DrVideo first transforms a long vi
deo into a coarse text-based long document to initially retrieve key frames and
then updates the documents with the augmented key frame information. It then emp
loys an agent-based iterative loop to continuously search for missing informatio

n and augment the document until sufficient question-related information is gathered for making the final predictions in a chain-of-thought manner. Extensive experiments on long video benchmarks confirm the effectiveness of our method. DrVideo significantly outperforms existing LLM-based state-of-the-art methods on EgoSchema benchmark (3 minutes), MovieChat-1K benchmark (10 minutes), and the long split of Video-MME benchmark (average of 44 minutes). Code is available at https://github.com/Upper9527/DrVideo.

****

Infighting in the Dark: Multi-Label Backdoor Attack in Federated Learning
Ye Li, Yanchao Zhao, Chengcheng Zhu, Jiale Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25770-25779
Federated Learning (FL), a privacy-preserving decentralized machine learning framework, has been shown to be vulnerable to backdoor attacks. Current research primarily focuses on the Single-Label Backdoor Attack (SBA), wherein adversaries share a consistent target. However, a critical fact is overlooked: adversaries may be non-cooperative, have distinct targets, and operate independently, which exhibits a more practical scenario called Multi-Label Backdoor Attack (MBA). Unfortunately, prior works are ineffective in the MBA scenario since non-cooperative attackers exclude each other. In this work, we conduct an in-depth investigation to uncover the inherent constraints of the exclusion: similar backdoor mappings are constructed for different targets, resulting in conflicts among backdoor functions. To address this limitation, we propose Mirage, the first non-cooperative MBA strategy in FL that allows attackers to inject effective and persistent backdoors into the global model without collusion by constructing in-distribution (ID) backdoor mapping. Specifically, we introduce an adversarial adaptation method to bridge the backdoor features and the target distribution in an ID manner. Additionally, we further leverage a constrained optimization method to ensure the ID mapping survives in the global training dynamics. Extensive evaluations demonstrate that Mirage outperforms various state-of-the-art attacks and bypasses existing defenses, achieving an average ASR greater than 97% and maintaining over 90% after 900 rounds. This work aims to alert researchers to this potential threat and inspire the design of effective defense mechanisms. Code has been made open-source.

****

Buffer Anytime: Zero-Shot Video Depth and Normal from Image Priors
Zhengfei Kuang, Tianyuan Zhang, Kai Zhang, Hao Tan, Sai Bi, Yiwei Hu, Zexiang Xu, Milos Hasan, Gordon Wetzstein, Fujun Luan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17660-17670
We present Buffer Anytime, a framework for estimation of depth and normal maps (which we call geometric buffers) from video that eliminates the need for paired video--depth and video--normal training data. Instead of relying on large-scale annotated video datasets, we demonstrate high-quality video buffer estimation by leveraging single-image priors with temporal consistency constraints. Our zero-shot training strategy combines state-of-the-art image estimation models based on optical flow smoothness through a hybrid loss function, implemented via a lightweight temporal attention architecture. Applied to leading image models like Depth Anything V2 and Marigold-E2E-FT, our approach significantly improves temporal consistency while maintaining accuracy. Experiments show that our method not only outperforms image-based approaches but also achieves results comparable to state-of-the-art video models trained on large-scale paired video datasets, despite using no such paired video data.

****

PSHuman: Photorealistic Single-image 3D Human Reconstruction using Cross-Scale Multiview Diffusion and Explicit Remeshing
Peng Li, Wangguandong Zheng, Yuan Liu, Tao Yu, Yangguang Li, Xingqun Qi, Xiaowei Chi, Siyu Xia, Yan-Pei Cao, Wei Xue, Wenhan Luo, Yike Guo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16008-16018
Photorealistic 3D human modeling is essential for various applications and has seen tremendous progress. However, existing methods for monocular full-body reconstruction, typically relying on front and/or predicted back view, still struggle

with satisfactory performance due to the ill-posed nature of the problem and so phisticated self-occlusions. In this paper, we propose PSHuman, a novel framewor k that explicitly reconstructs human meshes utilizing priors from the multiview diffusion model. It is found that directly applying multiview diffusion on singl e-view human images leads to severe geometric distortions, especially on genera ted faces. To address it, we propose a cross-scale diffusion that models the joi nt probability distribution of global full-body shape and local facial character istics, enabling identity-preserved novel-view generation without geometric dist ortion. Moreover, to enhance cross-view body shape consistency of varied human p oses, we condition the generative model on parametric models (SMPL-X), which pro vide body priors and prevent unnatural views inconsistent with human anatomy. Le veraging the generated multiview normal and color images, we present SMPLX-initi alized explicit human carving to recover realistic textured human meshes efficie ntly. Extensive experiments on CAPE and THuman2.1 demonstrate PSHuman's superior ity in geometry details, texture fidelity, and generalization capability.

****

LSNet: See Large, Focus Small

Ao Wang, Hui Chen, Zijia Lin, Jungong Han, Guiguang Ding; Proceedings of the Com puter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9718-9729

Vision network designs, including Convolutional Neural Networks and Vision Trans formers, have significantly advanced the field of computer vision. Yet, their co mplex computations pose challenges for practical deployments, particularly in re al-time applications. To tackle this issue, researchers have explored various li ghtweight and efficient network designs. However, existing lightweight models pr edominantly leverage self-attention mechanisms and convolutions for token mixing . This dependence brings limitations in effectiveness and efficiency in the perc eption and aggregation processes of lightweight networks, hindering the balance between performance and efficiency under limited computational budgets. In this paper, we draw inspiration from the dynamic heteroscale vision ability inherent in the efficient human vision system and propose a "See Large, Focus Small" stra tegy for lightweight vision network design. We introduce LS (Large-Small) convol ution, which combines large-kernel perception and small-kernel aggregation. It c an efficiently capture a wide range of perceptual information and achieve precis e feature aggregation for dynamic and complex visual representations, thus enabl ing proficient processing of visual information. Based on LS convolution, we pre sent LSNet, a new family of lightweight models. Extensive experiments demonstrat e that LSNet achieves superior performance and efficiency over existing lightwei ght networks in various vision tasks. Codes and models are available at https:// github.com/jameslahm/lsnet.

****

DynamicScaler: Seamless and Scalable Video Generation for Panoramic Scenes

Jinxiu Liu, Shaoheng Lin, Yinxiao Li, Ming-Hsuan Yang; Proceedings of the Comput er Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6144-6153

The increasing demand for immersive AR/VR applications and spatial intelligence has heightened the need to generate high-quality scene-level and 360deg panorami c video. However, most video diffusion models are constrained by limited resolut ion and aspect ratio, which restricts their applicability to scene-level dynamic content synthesis. In this work, we propose DynamicScaler, addressing these cha llenges by enabling spatially scalable and panoramic dynamic scene synthesis tha t preserves coherence across panoramic scenes of arbitrary size. Specifically, w e introduce a Offset Shifting Denoiser, facilitating efficient, synchronous, and coherent denoising panoramic dynamic scenes via a diffusion model with fixed re solution through a seamless rotating Window, which ensures seamless boundary tra nsitions and consistency across the entire panoramic space, accommodating varyin g resolutions and aspect ratios. Additionally, we employ a Global Motion Guidanc e mechanism to ensure both local detail fidelity and global motion continuity. E xtensive experiments demonstrate our method achieves superior content and motion quality in panoramic scene-level video generation, offering a training-free, ef ficient, and scalable solution for immersive dynamic scene creation with constan t VRAM consumption regardless of the output video resolution. Project page is av

ailable at https://dynamic-scaler.pages.dev/new.
****
Tartan IMU: A Light Foundation Model for Inertial Positioning in Robotics
Shibo Zhao, Sifan Zhou, Raphael Blanchard, Yuheng Qiu, Wenshan Wang, Sebastian S
cherer; Proceedings of the Computer Vision and Pattern Recognition Conference (C
VPR), 2025, pp. 22520-22529
Despite recent advances in deep learning, most existing learning IMU odometry me
thods are trained on specific datasets, lack generalization, and are prone to ov
erfitting, which limits their real-world application. To address these challenge
s, we present Tartan IMU, a foundation model designed for generalizable, IMU-bas
ed state estimation across diverse robotic platforms. Our approach consists of t
hree-stage: First, a pre-trained foundation model leverages over 100 hours of mu
lti-platform data to establish general motion knowledge, achieving 36% improveme
nt in ATE over specialized models. Second, to adapt to previously unseen tasks,
we employ the Low-Rank Adaptation (LoRA), allowing positive transfer with only 1
.1 M trainable parameters. Finally, to support robotics deployment, we introduce
 online test-time adaptation, which eliminates the boundary between training and
 testing, allowing the model to continuously "learn as it operates" at 200 FPS i
n real-time.
****
Event Ellipsometer: Event-based Mueller-Matrix Video Imaging
Ryota Maeda, Yunseong Moon, Seung-Hwan Baek; Proceedings of the Computer Vision
and Pattern Recognition Conference (CVPR), 2025, pp. 21804-21813
Light-matter interactions modify both the intensity and polarization state of li
ght. Changes in polarization, represented by a Mueller matrix, encode detailed s
cene information. Existing optical ellipsometers capture Mueller-matrix images;
however, they are often limited to static scenes due to long acquisition times.
Here, we introduce Event Ellipsometer, a method for acquiring Mueller-matrix ima
ges of dynamic scenes. Our imaging system employs fast-rotating quarter-wave pla
tes (QWPs) in front of a light source and an event camera that asynchronously ca
ptures intensity changes induced by the rotating QWPs. We develop an ellipsometr
ic-event image formation model, a calibration method, and an ellipsometric-event
 reconstruction method. We experimentally demonstrate that Event Ellipsometer en
ables Mueller-matrix imaging at 30fps, extending ellipsometry to dynamic scenes.
****
DocLayLLM: An Efficient Multi-modal Extension of Large Language Models for Text-
rich Document Understanding
Wenhui Liao, Jiapeng Wang, Hongliang Li, Chengyu Wang, Jun Huang, Lianwen Jin; P
roceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 202
5, pp. 4038-4049
Text-rich document understanding (TDU) requires comprehensive analysis of docume
nts containing substantial textual content and complex layouts. While Multimodal
 Large Language Models (MLLMs) have achieved fast progress in this domain, exist
ing approaches either demand significant computational resources or struggle wit
h effective multi-modal integration. In this paper, we introduce DocLayLLM, an e
fficient multi-modal extension of LLMs specifically designed for TDU. By lightly
 integrating visual patch tokens and 2D positional tokens into LLMs' input and e
ncoding the document content using the LLMs themselves, we fully take advantage
of the document comprehension capability of LLMs and enhance their perception of
 OCR information. We have also deeply considered the role of chain-of-thought (C
oT) and innovatively proposed the techniques of CoT Pre-training and CoT Anneali
ng. Our DocLayLLM can achieve remarkable performances with lightweight training
settings, showcasing its efficiency and effectiveness. Experimental results demo
nstrate that our DocLayLLM outperforms existing OCR-dependent methods and OCR-fr
ee competitors. Code and model are available at https://github.com/whlscut/DocLa
yLLM.
****
EDEN: Enhanced Diffusion for High-quality Large-motion Video Frame Interpolation
Zihao Zhang, Haoran Chen, Haoyu Zhao, Guansong Lu, Yanwei Fu, Hang Xu, Zuxuan Wu
; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR),

2025, pp. 2105-2115
Handling complex or nonlinear motion patterns has long posed challenges for video frame interpolation. Although recent advances in diffusion-based methods offer improvements over traditional optical flow-based approaches, they still struggle to generate sharp, temporally consistent frames in scenarios with large motion. To address this limitation, we introduce EDEN, an Enhanced Diffusion for high-quality large-motion vidEo frame iNterpolation. Our approach first utilizes a transformer-based tokenizer to produce refined latent representations of the intermediate frames for diffusion models. We then enhance the diffusion transformer with temporal attention across the process and incorporate a start-end frame difference embedding to guide the generation of dynamic motion. Extensive experiments demonstrate that EDEN achieves state-of-the-art results across popular benchmarks, including nearly a 10% LPIPS reduction on DAVIS and SNU-FILM, and an 8% improvement on DAIN-HD.
****

Handling Spatial-Temporal Data Heterogeneity for Federated Continual Learning via Tail Anchor
Hao Yu, Xin Yang, Le Zhang, Hanlin Gu, Tianrui Li, Lixin Fan, Qiang Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4874-4883
Federated Continual Learning (FCL) allows each client to continually update its knowledge from task streams, enhancing the applicability of federated learning in real-world scenarios. However, FCL needs to address not only spatial data heterogeneity between clients but also temporal data heterogeneity between tasks. In this paper, empirical experiments demonstrate that such input-level heterogeneity significantly affects the model's internal parameters and outputs, leading to severe spatial-temporal catastrophic forgetting of local and previous knowledge. To this end, we propose Federated Tail Anchor (FedTA) to mix trainable Tail Anchor with the frozen output features to adjust their position in the feature space, thereby overcoming parameter-forgetting and output-forgetting. Three novel components are also included: Input Enhancement for improving the performance of pre-trained models on downstream tasks; Selective Input Knowledge Fusion for fusion of heterogeneous local knowledge on the server; and Best Global Prototype Selection for finding the best anchor point for each class in the feature space. Extensive experiments demonstrate that FedTA not only outperforms existing FCL methods but also effectively preserves the relative positions of features.
****

DeSiRe-GS: 4D Street Gaussians for Static-Dynamic Decomposition and Surface Reconstruction for Urban Driving Scenes
Chensheng Peng, Chengwei Zhang, Yixiao Wang, Chenfeng Xu, Yichen Xie, Wenzhao Zheng, Kurt Keutzer, Masayoshi Tomizuka, Wei Zhan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6782-6791
We present DeSiRe-GS, a self-supervised gaussian splatting representation, enabling effective static-dynamic decomposition and high-fidelity surface reconstruction in complex driving scenarios. Our approach employs a two-stage optimization pipeline of dynamic street Gaussians. In the first stage, we extract 2D motion masks based on the observation that 3D Gaussian Splatting inherently can reconstruct only the static regions in dynamic environments. These extracted 2D motion priors are then mapped into the Gaussian space in a differentiable manner, leveraging an efficient formulation of dynamic Gaussians in the second stage. Combined with the introduced geometric regularizations, our method are able to address the over-fitting issues caused by data sparsity in autonomous driving, reconstructing physically plausible Gaussians that align with object surfaces rather than floating in air. Furthermore, we introduce temporal cross-view consistency to ensure coherence across time and viewpoints, resulting in high-quality surface reconstruction. Comprehensive experiments demonstrate the efficiency and effectiveness of DeSiRe-GS, surpassing prior self-supervised arts and achieving accuracy comparable to methods relying on external 3D bounding box annotations.
****

End-to-End HOI Reconstruction Transformer with Graph-based Encoding

Zhenrong Wang, Qi Zheng, Sihan Ma, Maosheng Ye, Yibing Zhan, Dongjiang Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27706-27715

Human-object interaction (HOI) reconstruction has garnered significant attention due to its diverse applications and the success of capturing human meshes. Existing HOI reconstruction methods often rely on explicitly modeling interactions between humans and objects. However, such a way leads to a natural conflict between 3D mesh reconstruction, which emphasizes global structure, and fine-grained contact reconstruction, which focuses on local details. To address the limitations of explicit modeling, we propose the End-to-End HOI Reconstruction Transformer with Graph-based Encoding (HOI-TG). It implicitly learns the interaction between humans and objects by leveraging self-attention mechanisms. Within the transformer architecture, we devise graph residual blocks to aggregate the topology among vertices of different spatial structures. This dual focus effectively balances global and local representations. Without bells and whistles, HOI-TG achieves state-of-the-art performance on BEHAVE and InterCap datasets. Particularly on the challenging InterCap dataset, our method improves the reconstruction results for human and object meshes by 8.9% and 8.6%, respectively.

****

REWIND: Real-Time Egocentric Whole-Body Motion Diffusion with Exemplar-Based Identity Conditioning

Jihyun Lee, Weipeng Xu, Alexander Richard, Shih-En Wei, Shunsuke Saito, Shaojie Bai, Te-Li Wang, Minhyuk Sung, Tae-Kyun Kim, Jason Saragih; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7095-7104

We present REWIND (Real-Time Egocentric Whole-Body Motion Diffusion), a one-step diffusion model for real-time, high-fidelity human motion estimation from egocentric image inputs. While an existing method for egocentric whole-body (i.e., body and hands) motion estimation is non-real-time and acausal due to diffusion-based iterative motion refinement to capture correlations between body and hand poses, REWIND operates in a fully causal and real-time manner. To enable real-time inference, we introduce (1) cascaded body-hand denoising diffusion, which effectively models the correlation between egocentric body and hand motions in a fast, feed-forward manner, and (2) diffusion distillation, which enables high-quality motion estimation with a single denoising step. Our denoising diffusion model is based on a modified Transformer architecture, designed to causally model output motions while enhancing generalizability to unseen motion lengths. Additionally, REWIND optionally supports identity-conditioned motion estimation when identity prior is available. To this end, we propose a novel identity conditioning method based on a small set of pose exemplars of the target identity, which further enhances motion estimation quality. Through extensive experiments, we demonstrate that REWIND significantly outperforms the existing baselines both with and without exemplar-based identity conditioning.

****

Hiding Images in Diffusion Models by Editing Learned Score Functions

Haoyu Chen, Yunqiao Yang, Nan Zhong, Kede Ma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18663-18673

Hiding data using neural networks (i.e., neural steganography) has achieved remarkable success across both discriminative classifiers and generative adversarial networks. However, the potential of data hiding in diffusion models remains relatively unexplored. Current methods exhibit limitations in achieving high extraction accuracy, model fidelity, and hiding efficiency due primarily to the entanglement of the hiding and extraction processes with multiple denoising diffusion steps. To address these, we describe a simple yet effective approach that embeds images at specific timesteps in the reverse diffusion process by editing the learned score functions. Additionally, we introduce a parameter-efficient fine-tuning method that combines gradient-based parameter selection with low-rank adaptation to enhance model fidelity and hiding efficiency. Comprehensive experiments demonstrate that our method extracts high-quality images at human-indistinguishable levels, replicates the original model behaviors at both sample and population levels, and embeds images orders of magnitude faster than prior methods. Besid

es, our method naturally supports multi-recipient scenarios through independent extraction channels.
****

Disco4D: Disentangled 4D Human Generation and Animation from a Single Image
Hui En Pang, Shuai Liu, Zhongang Cai, Lei Yang, Tianwei Zhang, Ziwei Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26331-26344
We present Disco4D, a novel Gaussian Splatting framework for 4D human generation and animation from a single image. Different from existing methods, Disco4D distinctively disentangles clothings (with Gaussian models) from the human body (with SMPL-X model), significantly enhancing the generation details and flexibility. It has the following technical innovations. (1) Disco4D learns to efficiently fit the clothing Gaussians over the SMPL-X Gaussians. (2) It adopts diffusion models to enhance the 3D generation process, e.g. modeling occluded parts not visible in the input image. (3) It learns an identity encoding for each clothing Gaussian to facilitate the separation and extraction of clothing assets. Furthermore, Disco4D naturally supports 4D human animation with vivid dynamics. Extensive experiments demonstrate the superiority of Disco4D on 4D human generation and animation tasks.
****

DoraCycle: Domain-Oriented Adaptation of Unified Generative Model in Multimodal Cycles
Rui Zhao, Weijia Mao, Mike Zheng Shou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2835-2846
Adapting generative models to specific domains presents an effective solution for satisfying specialized requirements. However, adapting to some complex domains remains challenging, especially when these domains require substantial paired data to capture the targeted distributions. Since unpaired data from a single modality, such as vision or language, is more readily available, we utilize the bidirectional mappings between vision and language learned by the unified generative model to enable training on unpaired data for domain adaptation. Specifically, we propose DoraCycle, which integrates two multimodal cycles: text-to-image-to-text and image-to-text-to-image. The model is optimized through cross-entropy loss computed at the cycle endpoints, where both endpoints share the same modality. This facilitates self-evolution of the model without reliance on annotated text-image pairs. Experimental results demonstrate that for tasks independent of paired knowledge, such as stylization, DoraCycle can effectively adapt the unified model using only unpaired data. For tasks involving new paired knowledge, such as specific identities, a combination of a small set of paired image-text examples and larger-scale unpaired data is sufficient for effective domain-oriented adaptation. The code will be released at https://github.com/showlab/DoraCycle.
****

WeatherGen: A Unified Diverse Weather Generator for LiDAR Point Clouds via Spider Mamba Diffusion
Yang Wu, Yun Zhu, Kaihua Zhang, Jianjun Qian, Jin Xie, Jian Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17019-17028
3D scene perception demands a large amount of adverse-weather LiDAR data, yet the cost of LiDAR data collection presents a significant scaling-up challenge. To this end, a series of LiDAR simulators have been proposed. Yet, they can only simulate a single adverse weather with a single physical model, and the fidelity is quite limited. This paper presents **WeatherGen**, the first unified diverse-weather LiDAR data diffusion generation framework, significantly improving fidelity. Specifically, we first design a map-based data producer, which is capable of providing a vast amount of high-quality diverse-weather data for training purposes. Then, we utilize the diffusion-denoising paradigm to construct a diffusion model. Among them, we propose a spider mamba generator with the spider mamba scan to restore the disturbed diverse weather data gradually. The spider mamba models the feature interactions by scanning the LiDAR beam circle and central ray, excellently maintaining the physical structure of the LiDAR point cloud. Subsequ

ently, we design a latent domain aligner following the generator to transfer rea l-world knowledge. Afterward, we devise a contrastive learning-based controller, which equips weather control signals with compact semantic knowledge through la nguage supervision from CLIP, guiding the diffusion model in generating more dis criminative data. Finally, we fine-tune WeatherGen with small-scale real-world d ata to further enhance its performance. Extensive evaluations on KITTI-360 and S eeing Through Fog demonstrate the high generation quality of WeatherGen. Through WeatherGen, we construct the mini-weather dataset, promoting the performance of the downstream task under adverse weather conditions.
****

MUST: The First Dataset and Unified Framework for Multispectral UAV Single Objec t Tracking
Haolin Qin, Tingfa Xu, Tianhao Li, Zhenxiang Chen, Tao Feng, Jianan Li; Proceedi ngs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16882-16891
UAV tracking faces significant challenges in real-world scenarios, such as small -size targets and occlusions, which limit the performance of RGB-based trackers. Multispectral images (MSI), which capture additional spectral information, offe r a promising solution to these challenges. However, progress in this field has been hindered by the lack of relevant datasets. To address this gap, we introduc e the first large-scale Multispectral UAV Single Object Tracking dataset (MUST), which includes 250 video sequences spanning diverse environments and challenges , providing a comprehensive data foundation for multispectral UAV tracking. We a lso propose a novel tracking framework, UNTrack, which encodes unified spectral, spatial, and temporal features from spectrum prompts, initial templates, and se quential searches. UNTrack employs an asymmetric transformer with a spectral bac kground eliminate mechanism for optimal relationship modeling and an encoder tha t continuously updates the spectrum prompt to refine tracking, improving both ac curacy and efficiency. Extensive experiments show that our proposed UNTrack outp erforms state-of-the-art UAV trackers. We believe our dataset and framework will drive future research in this area. The dataset is available on https://github. com/q2479036243/MUST-Multispectral-UAV-Single-Object-Tracking.
****

IDOL: Instant Photorealistic 3D Human Creation from a Single Image
Yiyu Zhuang, Jiaxi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, Wei Liu; Proceedings of the Computer Vision and Pattern Rec ognition Conference (CVPR), 2025, pp. 26308-26319
Creating a high-fidelity, animatable 3D full-body avatar from a single image is a challenging task due to the diverse appearance and poses of humans and the lim ited availability of high-quality training data. To achieve fast and high-qualit y human reconstruction, this work rethinks the task from the perspectives of dat aset, model, and representation. First, we introduce a large-scale HUman GEnerat ed training dataset, HuGe100K, consisting of 100K diverse, photorealistic human images with corresponding 24-view in a static pose or dynamic pose frames genera ted via a pose-controllable image-to-video model. Next, leveraging the diversity in views, poses, and appearances within HuGe100K, we develop a scalable feed-fo rward transformer model to predict a 3D human Gaussian representation in a unifo rm space of a given human image. This model is trained to disentangle human pose , shape, clothing geometry, and texture. Accordingly, the estimated Gaussians ca n be animated robustly without post-processing. We conduct comprehensive experim ents to validate the effectiveness of the proposed dataset and method. Our model demonstrates the generalizable ability to efficiently reconstruct photorealisti c humans in under 1 second using a single GPU. Additionally, it seamlessly suppo rts various applications, including animation, shape, and texture editing tasks.
****

Tightening Robustness Verification of MaxPool-based Neural Networks via Minimizi ng the Over-Approximation Zone
Yuan Xiao, Yuchen Chen, Shiqing Ma, Chunrong Fang, Tongtong Bai, Mingzheng Gu, Y uxin Cheng, Yanwei Chen, Zhenyu Chen; Proceedings of the Computer Vision and Pat tern Recognition Conference (CVPR), 2025, pp. 20695-20705

The robustness of neural network classifiers is important in the safety-critical domain and can be quantified by robustness verification. At present, efficient and scalable verification techniques are always sound but incomplete, and thus, the improvement of verified robustness results is the key criterion to evaluate the performance of incomplete verification approaches. The multi-variate function MaxPool is widely adopted yet challenging to verify. In this paper, we present Ti-Lin, a robustness verifier for MaxPool-based CNNs with Tight Linear Approximation. Following the sequel of minimizing the over-approximation zone of the non-linear function of CNNs, we are the first to propose the provably neuron-wise tightest linear bounds for the MaxPool function. By our proposed linear bounds, we can certify larger robustness results for CNNs. We evaluate the effectiveness of Ti-Lin on different verification frameworks with open-sourced benchmarks, including LeNet, PointNet, and networks trained on the MNIST, CIFAR-10, Tiny ImageNet and ModelNet40 datasets. Experimental results show that Ti-Lin significantly outperforms the state-of-the-art methods across all networks with up to 78.6% improvement in terms of the certified accuracy with almost the same time consumption as the fastest tool. Our code is available at https://anonymous.4open.science/r/Ti-Lin-cvpr-72EE.

****

SketchVideo: Sketch-based Video Generation and Editing

Feng-Lin Liu, Hongbo Fu, Xintao Wang, Weicai Ye, Pengfei Wan, Di Zhang, Lin Gao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23379-23390

Video generation and editing conditioned on text prompts or images have undergone significant advancements. However, challenges remain in accurately controlling global layout and geometry details solely by texts, and supporting motion control and local modification through images. In this paper, we aim to achieve sketch-based spatial and motion control for video generation and support fine-grained editing of real or synthetic videos. Based on the DiT video generation model, we propose a memory-efficient control structure with sketch control blocks that predict residual features of skipped DiT blocks. Sketches are drawn on one or two keyframes (at arbitrary time points) for easy interaction. To propagate such temporally sparse sketch conditions across all frames, we propose an inter-frame attention mechanism to analyze the relationship between the keyframes and each video frame. For sketch-based video editing, we design an additional video insertion module that maintains consistency between the newly edited content and the original video's spatial feature and dynamic motion. During inference, we use latent fusion for the accurate preservation of unedited regions. Extensive experiments demonstrate that our SketchVideo achieves superior performance in controllable video generation and editing.

****

PhysicsGen: Can Generative Models Learn from Images to Predict Complex Physical Relations?

Martin Spitznagel, Jan Vaillant, Janis Keuper; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11125-11134

The image-to-image translation abilities of generative learning models have recently made significant progress in the estimation of complex (steered) mappings between image distributions. While appearance based tasks like image in-painting or style transfer have been studied at length, we propose to investigate the potential of generative models in the context of physical simulations. Providing a dataset of 300k image-pairs and baseline evaluations for three different physical simulation tasks, we propose a benchmark to investigate the following research questions: i) are generative models able to learn complex physical relations from input-output image pairs? ii) what speedups can be achieved by replacing differential equation based simulations? While baseline evaluations of different current models show the potential for high speedups (ii), these results also show strong limitations toward the physical correctness (i). This underlines the need for new methods to enforce physical correctness.

****

Taste More, Taste Better: Diverse Data and Strong Model Boost Semi-Supervised Cr

owd Counting
Maochen Yang, Zekun Li, Jian Zhang, Lei Qi, Yinghuan Shi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24440-24451
Semi-supervised crowd counting is crucial for addressing the high annotation costs of densely populated scenes. Although several methods based on pseudo-labeling have been proposed, it remains challenging to effectively and accurately utilize unlabeled data. In this paper, we propose a novel framework called Taste More Taste Better (TMTB), which emphasizes both data and model aspects. Firstly, we explore a data augmentation technique well-suited for the crowd counting task. By inpainting the background regions, this technique can effectively enhance data diversity while preserving the fidelity of the entire scenes. Secondly, we introduce the Visual State Space Model as backbone to capture the global context information from crowd scenes, which is crucial for extremely crowded, low-light, and adverse weather scenarios. In addition to the traditional regression head for exact prediction, we employ an Anti-Noise classification head to provide less exact but more accurate supervision, since the regression head is sensitive to noise in manual annotations. We conduct extensive experiments on four benchmark datasets and show that our method outperforms state-of-the-art methods by a large margin. Code is publicly available on https://github.com/syhien/taste_more_taste_better.
****
Gaussian Splashing: Unified Particles for Versatile Motion Synthesis and Rendering
Yutao Feng, Xiang Feng, Yintong Shang, Ying Jiang, Chang Yu, Zeshun Zong, Tianjia Shao, Hongzhi Wu, Kun Zhou, Chenfanfu Jiang, Yin Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 518-529
We demonstrate the feasibility of integrating physics-based animations of solids and fluids with 3D Gaussian Splatting (3DGS) to create novel effects in virtual scenes reconstructed using 3DGS. Leveraging the coherence of the Gaussian Splatting and Position-Based Dynamics (PBD) in the underlying representation, we manage rendering, view synthesis, and the dynamics of solids and fluids in a cohesive manner. Similar to GaussianShader, we enhance each Gaussian kernel with an added normal, aligning the kernel's orientation with the surface normal to refine the PBD simulation. This approach effectively eliminates spiky noises that arise from rotational deformation in solids. It also allows us to integrate physically based rendering to augment the dynamic surface reflections on fluids. Consequently, our framework is capable of realistically reproducing surface highlights on dynamic fluids and facilitating interactions between scene objects and fluids from new views.
****
Improve Representation for Imbalanced Regression through Geometric Constraints
Zijian Dong, Yilei Wu, Chongyao Chen, Yingtian Zou, Yichi Zhang, Juan Helen Zhou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5082-5091
In representation learning, uniformity refers to the uniform feature distribution in the latent space (i.e., unit hypersphere). Previous work has shown that improving uniformity contributes to the learning of under-represented classes. However, most of the previous work focused on classification; the representation space of imbalanced regression remains unexplored. Classification-based methods are not suitable for regression tasks because they cluster features into distinct groups without considering the continuous and ordered nature essential for regression. In a geometric aspect, we uniquely focus on ensuring uniformity in the latent space for imbalanced regression through two key losses: enveloping and homogeneity. The enveloping loss encourages the induced trace to uniformly occupy the surface of a hypersphere, while the homogeneity loss ensures smoothness, with representations evenly spaced at consistent intervals. Our method integrates these geometric principles into the data representations via a Surrogate-driven Representation Learning (SRL) framework. Experiments with real-world regression and operator learning tasks highlight the importance of uniformity in imbalanced regression and validate the efficacy of our geometry-based loss functions.

****
AnyDressing: Customizable Multi-Garment Virtual Dressing via Latent Diffusion Models
Xinghui Li, Qichao Sun, Pengze Zhang, Fulong Ye, Zhichao Liao, Wanquan Feng, Songtao Zhao, Qian He; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23723-23733
Recent advances in garment-centric image generation from text and image prompts based on diffusion models are impressive. However, existing methods lack support for various combinations of attire, and struggle to preserve the garment details while maintaining faithfulness to the text prompts, limiting their performance across diverse scenarios. In this paper, we focus on a new task, i.e., Multi-Garment Virtual Dressing, and we propose a novel AnyDressing method for customizing characters conditioned on any combination of garments and any personalized text prompts. AnyDressing primarily comprises two primary networks named GarmentsNet and DressingNet, which are respectively dedicated to extracting detailed clothing features and generating customized images. Specifically, we propose an efficient and scalable module called Garment-Specific Feature Extractor in GarmentsNet to individually encode garment textures in parallel. This design prevents garment confusion while ensuring network efficiency. Meanwhile, we design an adaptive Dressing-Attention mechanism and a novel Instance-Level Garment Localization Learning strategy in DressingNet to accurately inject multi-garment features into their corresponding regions. This approach efficiently integrates multi-garment texture cues into generated images and further enhances text-image consistency. Additionally, we introduce a Garment-Enhanced Texture Learning strategy to improve the fine-grained texture details of garments. Thanks to our well-craft design, AnyDressing can serve as a plug-in module to easily integrate with any community control extensions for diffusion models, improving the diversity and controllability of synthesized images. Extensive experiments show that AnyDressing achieves state-of-the-art results.
****
Spectral Informed Mamba for Robust Point Cloud Processing
Ali Bahri, Moslem Yazdanpanah, Mehrdad Noori, Sahar Dastani, Milad Cheraghalikhani, Gustavo Adolfo Vargas Hakim, David Osowiechi, Farzad Beizaee, Ismail Ben Ayed, Christian Desrosiers; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11799-11809
State Space Models (SSMs) have shown significant promise in Natural Language Processing (NLP) and, more recently, computer vision. This paper introduces a new methodology leveraging Mamba and Masked Autoencoder (MAE) networks for point cloud data in both supervised and self-supervised learning. We propose three key contributions to enhance Mamba's capability in processing complex point cloud structures. First, we exploit the spectrum of a graph Laplacian to capture patch connectivity, defining an isometry-invariant traversal order that is robust to viewpoints and better captures shape manifolds than traditional 3D grid-based traversals. Second, we adapt segmentation via a recursive patch partitioning strategy informed by Laplacian spectral components, allowing finer integration and segment analysis. Third, we address token placement in MAE for Mamba by restoring tokens to their original positions, which preserves essential order and improves learning. Extensive experiments demonstrate our approach's improvements in classification, segmentation, and few-shot tasks over state-of-the-art (SOTA) baselines.
****
Latent Space Imaging
Matheus Souza, Yidan Zheng, Kaizhang Kang, Yogeshwar Nath Mishra, Qiang Fu, Wolfgang Heidrich; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28295-28305
Digital imaging systems have traditionally relied on brute-force measurement and processing of pixels arranged on regular grids. In contrast, the human visual system performs significant data reduction from the large number of photoreceptors to the optic nerve, effectively encoding visual information into a low-bandwidth latent space representation optimized for brain processing. Inspired by this, we propose a similar approach to advance artificial vision systems. Latent Spac

e Imaging introduces a new paradigm that combines optics and software to encode image information directly into the semantically rich latent space of a generative model. This approach substantially reduces bandwidth and memory demands during image capture and enables a range of downstream tasks focused on the latent space.We validate this principle through an initial hardware prototype based on a single-pixel camera. By implementing an amplitude modulation scheme that encodes into the generative model's latent space, we achieve compression ratios ranging from 1:100 to 1:1000 during imaging, and up to 1:16384 for downstream applications. This approach leverages the model's intrinsic linear boundaries, demonstrating the potential of latent space imaging for highly efficient imaging hardware, adaptable future applications in high-speed imaging, and task-specific cameras with significantly reduced hardware complexity.
****

Balanced Direction from Multifarious Choices: Arithmetic Meta-Learning for Domain Generalization

Xiran Wang, Jian Zhang, Lei Qi, Yinghuan Shi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30577-30587

Domain generalization is proposed to address distribution shift, arising from statistical disparities between training source and unseen target domains. The widely used first-order meta-learning algorithms demonstrate strong performance for domain generalization by leveraging the gradient matching theory, which aims to establish balanced parameters across source domains to reduce overfitting to any particular domain. However, our analysis reveals that there are actually numerous directions to achieve gradient matching, with current methods representing just one possible path. These methods actually overlook another critical factor that the balanced parameters should be close to the centroid of optimal parameters of each source domain. To address this, we propose a simple yet effective arithmetic meta-learning with arithmetic-weighted gradients. This approach, while adhering to the principles of gradient matching, promotes a more precise balance by estimating the centroid between domain-specific optimal parameters. Experimental results validate the effectiveness of our strategy. Our code is available at https://github.com/zzwdx/ARITH.
****

Anatomical Consistency and Adaptive Prior-informed Transformation for Multi-contrast MR Image Synthesis via Diffusion Model

Yejee Shin, Yeeun Lee, Hanbyol Jang, Geonhui Son, Hyeongyu Kim, Dosik Hwang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30918-30927

Multi-contrast magnetic resonance (MR) images offer critical diagnostic information but are limited by long scan times and high cost. While diffusion models (DMs) excel in medical image synthesis, they often struggle to maintain anatomical consistency and utilize the diverse characteristics of multi-contrast MR images effectively. We propose APT, a unified diffusion model designed to generate accurate and anatomically consistent multi-contrast MR images. APT introduces a mutual information fusion module and an anatomical consistency loss to preserve critical anatomical structures across multiple contrast inputs. To enhance synthesis, APT incorporates a two-stage inference process: in the first stage, a prior codebook provides coarse anatomical structures by selecting appropriate guidance based on precomputed similarity mappings and Bezier curve transformations. The second stage applies iterative unrolling with weighted averaging to refine the initial output, enhancing fine anatomical details and ensuring structural consistency. This approach enables the preservation of both global structures and local details, resulting in realistic and diagnostically valuable synthesized images. Extensive experiments on public multi-contrast MR brain images demonstrate that our approach significantly outperforms state-of-the-art methods. The source codes are available at https://github.com/yejees/APT.
****

BlobGEN-Vid: Compositional Text-to-Video Generation with Blob Video Representations

Weixi Feng, Chao Liu, Sifei Liu, William Yang Wang, Arash Vahdat, Weili Nie; Pro

ceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12989-12998

Existing video generation models struggle to follow complex text prompts and synthesize multiple objects, raising the need for additional grounding input for improved controllability. In this work, we propose to decompose videos into visual primitives -- blob video representation, a general representation for controllable video generation. Based on blob conditions, we develop a blob-grounded video diffusion model named BlobGEN-Vid that allows users to control object motions and fine-grained object appearance. In particular, we introduce a masked 3D attention module that effectively improves regional consistency across frames. In addition, we introduce a learnable module to interpolate text embeddings so that users can control semantics in specific frames and obtain smooth object transitions. We show that our framework is model-agnostic and build BlobGEN-Vid based on both U-Net and DiT-based video diffusion models. Extensive experimental results show that BlobGEN-Vid achieves superior zero-shot video generation ability and state-of-the-art layout controllability on multiple benchmarks. When combined with a Large Language Model for layout planning, our framework even outperforms proprietary text-to-video generators regarding compositional accuracy.

****

D2SP: Dynamic Dual-Stage Purification Framework for Dual Noise Mitigation in Vision-based Affective Recognition.

Haoran Wang, Xinji Mai, Zeng Tao, Xuan Tong, Junxiong Lin, Yan Wang, Jiawen Yu, Shaoqi Yan, Ziheng Zhou, Wenqiang Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19218-19229

The current advancements in Dynamic Facial Expression Recognition (DFER) methods mainly focus on better capturing the spatial and temporal features of facial expressions. However, DFER datasets contain a substantial amount of noisy samples, and few have addressed the issue of handling this noise. We identified two types of noise: one is caused by low-quality data resulting from factors such as occlusion, dim lighting, and blurriness; the other arises from mislabeled data due to annotation bias by annotators. Addressing the two types of noise, we have meticulously crafted a Dynamic Dual-Stage Purification (D2SP) Framework. This initiative aims to dynamically purify the DFER datasets of these two types of noise, ensuring that only high-quality and correctly labeled data is used in the training process. To mitigate low-quality samples, we introduce the Coarse-Grained Pruning (CGP) stage, which computes sample weights and prunes those low-weight samples. After CGP, the Fine-Grained Correction (FGC) stage evaluates prediction stability to correct mislabeled data. Moreover, D2SP is conceived as a general and plug-and-play framework, tailored to integrate seamlessly with prevailing DFER methods. Extensive experiments covering prevalent DFER datasets and deploying multiple benchmark methods have substantiated D2SP's ability to significantly enhance performance metrics.

****

PartRM: Modeling Part-Level Dynamics with Large Cross-State Reconstruction Model

Mingju Gao, Yike Pan, Huan-ang Gao, Zongzheng Zhang, Wenyi Li, Hao Dong, Hao Tang, Li Yi, Hao Zhao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7004-7014

As interest grows in world models that predict future states from current observations and actions, accurately modeling part-level dynamics has become increasingly relevant for various applications. Existing approaches, such as Puppet-Master, rely on fine-tuning large-scale pre-trained video diffusion models, which are impractical for real-world use due to the limitations of 2D video representation and slow processing times. To overcome these challenges, we present PartRM, a novel 4D reconstruction framework that simultaneously models appearance, geometry, and part-level motion from multi-view images of a static object. PartRM builds upon large 3D Gaussian reconstruction models, leveraging their extensive knowledge of appearance and geometry in static objects. To address data scarcity in 4D, we introduce the PartDrag-4D dataset, providing multi-view observations of part-level dynamics across over 20,000 states. We enhance the model's understanding of interaction conditions with a multi-scale drag embedding module that captur

es dynamics at varying granularities. To prevent catastrophic forgetting during fine-tuning, we implement a two-stage training process that focuses sequentially on motion and appearance learning. Experimental results show that PartRM establishes a new state-of-the-art in part-level motion learning and can be applied in manipulation tasks in robotics. Our code, data, and models are publicly available to facilitate future research.

****

LaVin-DiT: Large Vision Diffusion Transformer

Zhaoqing Wang, Xiaobo Xia, Runnan Chen, Dongdong Yu, Changhu Wang, Mingming Gong, Tongliang Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20060-20070

This paper presents the Large Vision Diffusion Transformer (LaVin-DiT), a scalable and unified foundation model designed to tackle over 20 computer vision tasks in a generative framework. Unlike existing large vision models directly adapted from natural language processing architectures, which rely on less efficient autoregressive techniques and disrupt spatial relationships essential for vision data, LaVin-DiT introduces key innovations to optimize generative performance for vision tasks. First, to address the high dimensionality of visual data, we incorporate a spatial-temporal variational autoencoder that encodes data into a continuous latent space. Second, for generative modeling, we develop a joint diffusion transformer that progressively produces vision outputs. Third, for unified multi-task training, in-context learning is implemented. Input-target pairs serve as task context, which guides the diffusion transformer to align outputs with specific tasks within the latent space. During inference, a task-specific context set and test data as queries allow LaVin-DiT to generalize across tasks without fine-tuning. Trained on extensive vision datasets, the model is scaled from 0.1B to 3.4B parameters, demonstrating substantial scalability and state-of-the-art performance across diverse vision tasks. This work introduces a novel pathway for large vision foundation models, underscoring the promising potential of diffusion transformers. The code and models are available at https://derrickwang005.github.io/LaVin-DiT.

****

DiffFNO: Diffusion Fourier Neural Operator

Xiaoyi Liu, Hao Tang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 150-160

We introduce DiffFNO, a novel diffusion framework for arbitrary-scale super-resolution strengthened by a Weighted Fourier Neural Operator (WFNO). Mode Rebalancing in WFNO effectively captures critical frequency components, significantly improving the reconstruction of high-frequency image details that are crucial for super-resolution tasks. Gated Fusion Mechanism (GFM) adaptively complements WFNO's spectral features with spatial features from an Attention-based Neural Operator (AttnNO). This enhances the network's capability to capture both global structures and local details. Adaptive Time-Step (ATS) ODE solver, a deterministic sampling strategy, accelerates inference without sacrificing output quality by dynamically adjusting integration step sizes ATS. Extensive experiments demonstrate that DiffFNO achieves state-of-the-art (SOTA) results, outperforming existing methods across various scaling factors by a margin of 2-4 dB in PSNR, including those beyond the training distribution. It also achieves this at lower inference time. Our approach sets a new standard in super-resolution, delivering both superior accuracy and computational efficiency.

****

CAP-Net: A Unified Network for 6D Pose and Size Estimation of Categorical Articulated Parts from a Single RGB-D Image

Jingshun Huang, Haitao Lin, Tianyu Wang, Yanwei Fu, Xiangyang Xue, Yi Zhu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11654-11664

This paper tackles category-level pose estimation of ar- ticulated objects in robotic manipulation tasks and intro- duces a new benchmark dataset. While recent methods es- timate part poses and sizes at the category level, they often rely on geometric cues and complex multi-stage pipelines that first segment parts from

the point cloud, followed by Normalized Part Coordinate Space (NPCS) estimation for 6D poses. These approaches overlook dense semantic cues from RGB images, leading to suboptimal accuracy, partic- ularly for objects with small parts. To address these limita- tions, we propose a single-stage Network, CAP-Net, for es- timating the 6D poses and sizes of Categorical Articulated Parts. This method combines RGB-D features to generate instance segmentation and NPCS representations for each part in an end-to-end manner. CAP-Net uses a unified net- work to simultaneously predict point-wise class labels, cen- troid offsets, and NPCS maps. A clustering algorithm then groups points of the same predicted class based on their es- timated centroid distances to isolate each part. Finally, the NPCS region of each part is aligned with the point cloud to recover its final pose and size. To bridge the sim-to-real do- main gap, we introduce the RGBD-Art dataset, the largest RGB-D articulated dataset to date, featuring photorealistic RGB images and depth noise simulated from real sensors. Experimental evaluations on the RGBD-Art dataset demon- strate that our method significantly outperforms the state-of-the-art approach. Real-world deployments of our model in robotic tasks underscore its robustness and exceptional sim-to-real transfer capabilities, confirming its substantial practical utility.

****

SeCap: Self-Calibrating and Adaptive Prompts for Cross-view Person Re-Identification in Aerial-Ground Networks

Shining Wang, Yunlong Wang, Ruiqi Wu, Bingliang Jiao, Wenxuan Wang, Peng Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22119-22128

When discussing the Aerial-Ground Person Re-identification (AGPReID) task, we face the main challenge of the significant appearance variations caused by different viewpoints, making identity matching difficult. To address this issue, previous methods attempt to reduce the differences between viewpoints by critical attributes and decoupling the viewpoints. While these methods can mitigate viewpoint differences to some extent, they still face two main issues: (1) difficulty in handling viewpoint diversity and (2) neglect of the contribution of local features. To effectively address these challenges, we design and implement the Self-Calibrating and Adaptive Prompt (SeCap) method for the AGPReID task. The core of this framework relies on the Prompt Re-calibration Module (PRM), which adaptively re-calibrates prompts based on the input. Combined with the Local Feature Refinement Module (LFRM), SeCap can extract view-invariant features from local features for AGPReID. Meanwhile, given the current scarcity of datasets in the AGPReID field, we further contribute two real-world Large-scale Aerial-Ground Person Re-Identification datasets, LAGPeR and G2APS-ReID. The former is collected and annotated by us independently, covering 4,231 unique identities and containing 63,841 high-quality images; the latter is reconstructed from the person search dataset G2APS. Through extensive experiments on AGPReID datasets, we demonstrate that SeCap is a feasible and effective solution for the AGPReID task.

****

Zero-Shot Styled Text Image Generation, but Make It Autoregressive

Vittorio Pippi, Fabio Quattrini, Silvia Cascianelli, Alessio Tonioni, Rita Cucchiara; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7910-7919

Styled Handwritten Text Generation (HTG) has recently received attention from the computer vision and document analysis communities, which have developed several solutions, either GAN- or diffusion-based, that achieved promising results. Nonetheless, these strategies fail to generalize to novel styles and have technical constraints, particularly in terms of maximum output length and training efficiency. To overcome these limitations, in this work, we propose a novel framework for text-image generation, dubbed Emuru. Our approach leverages a powerful text-image representation model (a variational autoencoder) combined with an autoregressive Transformer. Our approach enables the generation of styled text images conditioned on textual content and style examples, such as specific fonts or handwriting styles. We train our model solely on a diverse, synthetic dataset of English text rendered in over 100,000 typewritten and calligraphy fonts, which give

s it the capability to reproduce unseen styles (both fonts and users' handwritin g) in zero-shot. To the best of our knowledge, Emuru is the first autoregressive model for HTG, and the first designed specifically for generalization to novel styles. Moreover, our model generates images without background artifacts, which are easier to use for downstream applications. Extensive evaluation on both typ ewritten and handwritten, any-length text image generation scenarios demonstrate s the effectiveness of our approach.
****

Don't Shake the Wheel: Momentum-Aware Planning in End-to-End Autonomous Driving
Ziying Song, Caiyan Jia, Lin Liu, Hongyu Pan, Yongchang Zhang, Junming Wang, Xin gyu Zhang, Shaoqing Xu, Lei Yang, Yadan Luo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22432-22441
End-to-end autonomous driving frameworks enable seamless integration of percepti on and planning but often rely on one-shot trajectory prediction, which may lead to unstable control and vulnerability to occlusions in single-frame perception. To address this, we propose the Momentum-Aware Driving (MomAD) framework, which introduces trajectory momentum and perception momentum to stabilize and refine trajectory predictions. MomAD comprises two core components: (1) Topological Tra jectory Matching (TTM) employs Hausdorff Distance to select the optimal planning query that aligns with prior paths to ensure coherence; (2) Momentum Planning I nteractor (MPI) cross-attends the selected planning query with historical querie s to expand static and dynamic perception files. This enriched query, in turn, h elps regenerate long-horizon trajectory and reduce collision risks. To mitigate noise arising from dynamic environments and detection errors, we introduce robus t instance denoising during training, enabling the planning model to focus on cr itical signals and improve its robustness. We also propose a novel Trajectory Pr ediction Consistency (TPC) metric to quantitatively assess planning stability. E xperiments on the nuScenes dataset demonstrate that MomAD achieves superior long -term consistency (>3s) compared to SOTA methods. Moreover, evaluations on the c urated Turning-nuScenes shows that MomAD reduces the collision rate by 26% and i mproves TPC by 0.97m (33.45%) over a 6s prediction horizon, while closed-loop on Bench2Drive demonstrates an up to 16.3% improvement in success rate.
****

Leveraging Perturbation Robustness to Enhance Out-of-Distribution Detection
Wenxi Chen, Raymond A. Yeh, Shaoshuai Mou, Yan Gu; Proceedings of the Computer V ision and Pattern Recognition Conference (CVPR), 2025, pp. 4724-4733
Out-of-distribution (OOD) detection is the task of identifying inputs that devia te from the training data distribution. This capability is essential for the saf e deployment of deep computer vision models in open-world environments. In this work, we propose a post-hoc method, Perturbation-Rectified OOD detection (PRO), based on the insight that prediction confidence for OOD inputs is more susceptib le to reduction under perturbation than IND inputs. From this observation, we pr oposed a meta-score function that searches for local minimum scores near origina l inputs by applying gradient descent. This procedure enhances the separability between in-distribution (IND) and OOD samples. Importantly, the approach improve s OOD detection performance without complex modifications to the underlying mode l architectures or training protocol. To validate our approach, we conduct exten sive experiments using the OpenOOD benchmark. Our approach further pushes the li mit of softmax-based OOD detection and is the leading post-hoc method for small- scale models. On a CIFAR-10 model with adversarial training, PRO effectively det ects near-OOD inputs, achieving a reduction of more than 10% on FPR@95 compared to state-of-the-art methods.
****

Neural Motion Simulator Pushing the Limit of World Models in Reinforcement Learn ing
Chenjie Hao, Weyl Lu, Yifan Xu, Yubei Chen; Proceedings of the Computer Vision a nd Pattern Recognition Conference (CVPR), 2025, pp. 27608-27617
An embodied system must not only model the patterns of the external world but al so understand its own motion dynamics. A motion dynamic model is essential for e fficient skill acquisition and effective planning. In this work, we introduce th

e neural motion simulator (MoSim), a world model that predicts the future physic al state of an embodied system based on current observations and actions. MoSim achieves state-of-the-art performance in physical state prediction and provides competitive performance across a range of downstream tasks. This works shows tha t when a world model is accurate enough and performs precise long-horizon predic tions, it can facilitate efficient skill acquisition in imagined worlds and even enable zero-shot reinforcement learning. Furthermore, MoSim can transform any m odel-free reinforcement learning (RL) algorithm into a model-based approach, eff ectively decoupling physical environment modeling from RL algorithm development. This separation allows for independent advancements in RL algorithms and world modeling, significantly improving sample efficiency and enhancing generalization capabilities. Our findings highlight that world models for motion dynamics is a promising direction for developing more versatile and capable embodied systems.
****

Aesthetic Post-Training Diffusion Models from Generic Preferences with Step-by-s tep Preference Optimization

Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Mingxi Cheng, J i Li, Liang Zheng; Proceedings of the Computer Vision and Pattern Recognition Co nference (CVPR), 2025, pp. 13199-13208

Generating visually appealing images is fundamental to modern text-to-image gene ration models. A potential solution to better aesthetics is direct preference op timization (DPO), which has been applied to diffusion models to improve general image quality including prompt alignment and aesthetics. Popular DPO methods pro pagate preference labels from clean image pairs to all the intermediate steps al ong the two generation trajectories. However, preference labels provided in exis ting datasets are blended with layout and aesthetic opinions, which would disagr ee with aesthetic preference. Even if aesthetic labels were provided (at substan tial cost), it would be hard for the two-trajectory methods to capture nuanced v isual differences at different steps. To improve aesthetics economically, this p aper uses existing generic preference data and introduces step-by-step preferenc e optimization (SPO) that discards the propagation strategy and allows fine-grai ned image details to be assessed. Specifically, at each denoising step, we 1) sa mple a pool of candidates by denoising from a shared noise latent, 2) use a step -aware preference model to find a suitable win-lose pair to supervise the diffus ion model, and 3) randomly select one from the pool to initialize the next denoi sing step. This strategy ensures that diffusion models focus on the subtle, fine -grained visual differences instead of layout aspect. We find that aesthetics ca n be significantly enhanced by accumulating these improved minor differences. Wh en fine-tuning Stable Diffusion v1.5 and SDXL, SPO yields significant improvemen ts in aesthetics compared with existing DPO methods while not sacrificing image-text alignment compared with vanilla models. Moreover, SPO converges much faster than DPO methods due to the use of more correct preference labels provided by t he step-aware preference model. Code and models are available at https://github. com/RockeyCoss/SPO.
****

Adversarial Diffusion Compression for Real-World Image Super-Resolution

Bin Chen, Gehui Li, Rongyuan Wu, Xindong Zhang, Jie Chen, Jian Zhang, Lei Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2 025, pp. 28208-28220

Real-world image super-resolution (Real-ISR) aims to reconstruct high-resolution images from low-resolution inputs degraded by complex, unknown processes. While many Stable Diffusion (SD)-based Real-ISR methods have achieved remarkable succ ess, their slow, multi-step inference hinders practical deployment. Recent SD-ba sed one-step networks like OSEDiff and S3Diff alleviate this issue but still inc ur high computational costs due to their reliance on large pretrained SD models. This paper proposes a novel Real-ISR method, AdcSR, by distilling the one-step diffusion network OSEDiff into a streamlined diffusion-GAN model under our Adver sarial Diffusion Compression (ADC) framework. We meticulously examine the module s of OSEDiff, categorizing them into two types: (1) Removable (VAE encoder, prom pt extractor, text encoder, etc.) and (2) Prunable (denoising UNet and VAE decod

er). Since direct removal and pruning can degrade the model's generation capabil ity, we pretrain our pruned VAE decoder to restore its ability to decode images and employ adversarial distillation to compensate for performance loss. This ADC -based diffusion-GAN hybrid design effectively reduces complexity by 73% in infe rence time, 78% in computation, and 74% in parameters, while preserving the mode l's generation capability. Experiments manifest that our proposed AdcSR achieves competitive recovery quality on both synthetic and real-world datasets, offerin g up to 9.3x speedup over previous one-step diffusion-based methods. Code and mo dels are available at https://github.com/Guaishou74851/AdcSR.

****

DiSciPLE: Learning Interpretable Programs for Scientific Visual Discovery

Utkarsh Mall, Cheng Perng Phoo, Mia Chiquier, Bharath Hariharan, Kavita Bala, Ca rl Vondrick; Proceedings of the Computer Vision and Pattern Recognition Conferen ce (CVPR), 2025, pp. 29258-29267

Visual data is used in numerous different scientific workflows ranging from remo te sensing to ecology. As the amount of observation data increases, the challeng e is not just to make accurate predictions but also to understand the underlying mechanisms for those predictions. Good interpretation is important in scientifi c workflows, as it allows for better decision-making by providing insights into the data. This paper introduces an automatic way of obtaining such interpretable -by-design models, by learning programs that interleave neural networks. We prop ose DiSciPLE (Discovering Scientific Programs using LLMs and Evolution) an evolu tionary algorithm that leverages common sense and prior knowledge of large langu age models (LLMs) to create Python programs explaining visual data. Additionally , we propose two improvements: a program critic and a program simplifier to impr ove our method further to synthesize good programs. On three different real worl d problems, DiSciPLE learns state-of-the-art programs on novel tasks with no pri or literature. For example, we can learn programs with 35% lower error than the closest non-interpretable baseline for population density estimation.

****

SOLAMI: Social Vision-Language-Action Modeling for Immersive Interaction with 3D Autonomous Characters

Jianping Jiang, Weiye Xiao, Zhengyu Lin, Huaizhong Zhang, Tianxiang Ren, Yang Ga o, Zhiqian Lin, Zhongang Cai, Lei Yang, Ziwei Liu; Proceedings of the Computer V ision and Pattern Recognition Conference (CVPR), 2025, pp. 26887-26898

Human beings are social animals. How to equip 3D autonomous characters with simi lar social intelligence that can perceive, understand and interact with humans r emains an open yet foundamental problem. In this paper, we introduce SOLAMI, the first end-to-end Social vision-Language-Action (VLA) Modeling framework for Imm ersive interaction with 3D autonomous characters. Specifically, SOLAMI builds 3D autonomous characters from three aspects: 1) Social VLA Architecture: We propos e a unified social VLA framework to generate multimodal response (speech and mot ion) based on the user's multimodal input to drive the character for social inte raction. 2) Interactive Multimodal Data: We present SynMSI, a synthetic multimo dal social interaction dataset generated by an automatic pipeline using only exi sting motion datasets to address the issue of data scarcity. 3) Immersive VR Int erface: We develop a VR interface that enables users to immersively interact wit h these characters driven by various architectures. Extensive quantitative exper iments and user studies demonstrate that our framework leads to more precise and natural character responses (in both speech and motion) that align with user ex pectations with lower latency.

****

EntropyMark: Towards More Harmless Backdoor Watermark via Entropy-based Constrai nt for Open-source Dataset Copyright Protection

Ming Sun, Rui Wang, Zixuan Zhu, Lihua Jing, Yuanfang Guo; Proceedings of the Com puter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30692-30701

High-quality open-source datasets are essential for advancing deep neural networ ks. However, the unauthorized commercial use of these datasets has raised signif icant concerns about copyright protection. One promising approach is backdoor wa termark-based dataset ownership verification (BW-DOV), in which dataset protecto

rs implant specific backdoors into illicit models through dataset watermarking, enabling the tracing of these models through abnormal prediction behaviors. Unfo rtunately, the targeted nature of these BW-DOV methods can be maliciously exploi ted, potentially leading to harmful side effects. While existing harmless method s attempt to mitigate these risks, watermarked datasets can still negatively aff ect prediction results, partially compromising dataset functionality. In this pa per, we propose a more harmless backdoor watermark, called EntropyMark, which im proves prediction confidence without altering the final prediction results. For this purpose, an entropy-based constraint is introduced to regulate the probabil ity distribution. Specifically, we design an iterative clean-label dataset water marking framework. Our framework employs gradient matching and adaptive data sel ection to optimize backdoor injection. In parallel, we introduce a hypothesis te st method grounded in entropy inconsistency to verify dataset ownership. Extensi ve experiments on benchmark datasets demonstrate the effectiveness, transferabil ity, and defense resistance of our approach.

****

Adaptive Markup Language Generation for Contextually-Grounded Visual Document Un derstanding

Han Xiao, Yina Xie, Guanxin Tan, Yinghao Chen, Rui Hu, Ke Wang, Aojun Zhou, Hao Li, Hao Shao, Xudong Lu, Peng Gao, Yafei Wen, Xiaoxin Chen, Shuai Ren, Hongsheng Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29558-29568

Visual Document Understanding has become essential with the increase of text-ric h visual content. This field poses significant challenges due to the need for ef fective integration of visual perception and textual comprehension, particularly across diverse document types with complex layouts. Moreover, existing fine-tun ing datasets for this domain often fall short in providing the detailed contextu al information for robust understanding, leading to hallucinations and limited c omprehension of spatial relationships among visual elements. To address these ch allenges, we propose an innovative pipeline that utilizes adaptive generation of markup languages, such as Markdown, JSON, HTML, and TiKZ, to build highly struc tured document representations and deliver contextually-grounded responses. We i ntroduce two fine-grained structured datasets: DocMark-Pile, comprising approxim ately 3.8M pretraining data pairs for document parsing, and DocMark-Instruct, fe aturing 624k fine-tuning data annotations for grounded instruction following.Ext ensive experiments demonstrate that our proposed model significantly outperforms existing state-of-the-art MLLMs across a range of visual document understanding benchmarks, facilitating advanced reasoning and comprehension capabilities in c omplex visual scenarios.

****

BARD-GS: Blur-Aware Reconstruction of Dynamic Scenes via Gaussian Splatting

Yiren Lu, Yunlai Zhou, Disheng Liu, Tuo Liang, Yu Yin; Proceedings of the Comput er Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16532-16542

3D Gaussian Splatting (3DGS) has shown remarkable potential for static scene rec onstruction, and recent advancements have extended its application to dynamic sc enes. However, the quality of reconstructions depends heavily on high-quality in put images and precise camera poses, which is not that trivial to fulfill in the real-world scenarios. Capturing dynamic scenes with handheld monocular cameras, for instance, typically involves simultaneous movement of both the camera and o bjects within a single exposure. This combined motion frequently results in imag e blur that existing methods cannot adequately handle. To address these challeng es, we introduce BARD-GS, a novel approach for robust dynamic scene reconstructi on that effectively handles blurry inputs and imprecise camera poses. Our method comprises two main components: 1) camera motion deblurring and 2) object motion deblurring. By explicitly decomposing motion blur into camera motion blur and o bject motion blur and modeling them separately, we achieve significantly improve d rendering results in dynamic regions. In addition, we collect a real-world mot ion blur dataset of dynamic scenes to evaluate our approach. Extensive experimen ts demonstrate that BARD-GS effectively reconstructs high-quality dynamic scenes under realistic conditions, significantly outperforming existing methods.

****
SALAD: Skeleton-aware Latent Diffusion for Text-driven Motion Generation and Editing
Seokhyeon Hong, Chaelin Kim, Serin Yoon, Junghyun Nam, Sihun Cha, Junyong Noh; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7158-7168
Text-driven motion generation has advanced significantly with the rise of denoising diffusion models. However, previous methods often oversimplify representations for the skeletal joints, temporal frames, and textual words, limiting their ability to fully capture the information within each modality and their interactions. Moreover, when using pre-trained models for downstream tasks, such as editing, they typically require additional efforts, including manual interventions, optimization, or fine-tuning. In this paper, we introduce a skeleton-aware latent diffusion (SALAD), a model that explicitly captures the intricate inter-relationships between joints, frames, and words. Furthermore, by leveraging cross-attention maps produced during the generation process, we enable the attention-based zero-shot text-driven motion editing using a pre-trained SALAD model, requiring no additional user input beyond text prompts. Our approach significantly outperforms previous methods in terms of text-motion alignment without compromising generation quality, and demonstrates practical versatility by providing diverse editing capabilities beyond generation. Code is available at project page.
****
Towards Universal AI-Generated Image Detection by Variational Information Bottleneck Network
Haifeng Zhang, Qinghui He, Xiuli Bi, Weisheng Li, Bo Liu, Bin Xiao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23828-23837
The rapid advancement of generative models has significantly improved the quality of generated images. Meanwhile, it challenges information authenticity and credibility. Current generated image detection methods based on large-scale pre-trained multimodal models have achieved impressive results. Although these models provide abundant features, the authentication task-related features are often submerged. Consequently, those authentication task-irrelated features cause models to learn superficial biases, thereby harming their generalization performance across different model genera (e.g., GANs and Diffusion Models). To this end, we proposed VIB-Net, which uses Variational Information Bottlenecks to enforce authentication task-related feature learning. We tested and analyzed the proposed method and existing methods on samples generated by 17 different generative models. Compared to SOTA methods, VIB-Net achieved a 5.55% improvement in mAP and a 9.33% increase in accuracy. Notably, in generalization tests on unseen generative models from different series, VIB-Net improved mAP by 12.48% and accuracy by 23.59% over SOTA methods. The code is available at https://github.com/oceanzhf/VIBAIGCDetect.
****
HSI: A Holistic Style Injector for Arbitrary Style Transfer
Shuhao Zhang, Hui Kang, Yang Liu, Fang Mei, Hongjuan Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23433-23442
Attention-based arbitrary style transfer methods have gained significant attention recently due to their impressive ability to synthesize style details. However, the point-wise matching within the attention mechanism may overly focus on local patterns such that neglect the remarkable global features of style images. Additionally, when processing large images, the quadratic complexity of the attention mechanism will bring high computational load. To alleviate above problems, we propose Holistic Style Injector (HSI), a novel attention-style transformation module to deliver artistic expression of target style. Specifically, HSI performs stylization only based on global style representation that is more in line with the characteristics of style transfer, to avoid generating local disharmonious patterns in stylized images. Moreover, we propose a dual relation learning mechanism inside the HSI to dynamically render images by leveraging semantic similarity in content and style, ensuring the stylized images preserve the original con

tent and improve style fidelity. Note that the proposed HSI achieves linear comp utational complexity because it establishes feature mapping through element-wise multiplication rather than matrix multiplication. Qualitative and quantitative results demonstrate that our method outperforms state-of-the-art approaches in b oth effectiveness and efficiency.
****

LookingGlass: Generative Anamorphoses via Laplacian Pyramid Warping
Pascal Chang, Sergio Sancho, Jingwei Tang, Markus Gross, Vinicius Azevedo; Proce edings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, p p. 24-33
Anamorphosis refers to a category of images that are intentionally distorted, ma king them unrecognizable when viewed directly. Their true form only reveals itse lf when seen from a specific viewpoint, which can be through some catadioptric d evice like a mirror or a lens. While the construction of these mathematical devi ces can be traced back to as early as the 17th century, they are only interpreta ble when viewed from a specific vantage point and tend to lose meaning when seen normally. In this paper, we revisit these famous optical illusions with a gener ative twist. With the help of latent rectified flow models, we propose a method to create anamorphic images that still retain a valid interpretation when viewed directly. To this end, we introduce Laplacian Pyramid Warping, a frequency-awar e image warping technique key to generating high-quality visuals. Our work exten ds Visual Anagrams (Geng et al. 2024) to latent space models and to a wider rang e of spatial transforms, enabling the creation of novel generative perceptual il lusions.
****

V2V3D: View-to-View Denoised 3D Reconstruction for Light Field Microscopy
Jiayin Zhao, Zhenqi Fu, Tao Yu, Hui Qiao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26451-26461
Light field microscopy (LFM) has gained significant attention due to its ability to capture snapshot-based, large-scale 3D fluorescence images. However, existin g LFM reconstruction algorithms are highly sensitive to sensor noise or require hard-to-get ground-truth annotated data for training. To address these challenge s, this paper introduces V2V3D, an unsupervised view2view-based framework that e stablishes a new paradigm for joint optimization of image denoising and 3D recon struction in a unified architecture. We assume that the LF images are derived fr om a consistent 3D signal, with the noise in each view being independent. This e nables V2V3D to incorporate the principle of noise2noise for effective denoising . To enhance the recovery of high-frequency details, we propose a novel wave-opt ics-based feature alignment technique, which transforms the point spread functio n, used for forward propagation in wave optics, into convolution kernels specifi cally designed for feature alignment. Moreover, we introduce an LFM dataset cont aining LF images and their corresponding 3D intensity volumes. Extensive experim ents demonstrate that our approach achieves high computational efficiency and ou tperforms the other state-of-the-art methods. These advancements position V2V3D as a promising solution for 3D imaging under challenging conditions. Our code an d dataset will be publicly accessible at https://joey1998hub.github.io/V2V3D/.
****

DiN: Diffusion Model for Robust Medical VQA with Semantic Noisy Labels
Erjian Guo, Zhen Zhao, Zicheng Wang, Tong Chen, Yunyi Liu, Luping Zhou; Proceedi ngs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14337-14346
Medical Visual Question Answering (Med-VQA) systems benefit the interpretation o f medical images containing critical clinical information. However, the challeng e of noisy labels and limited high-quality datasets remains underexplored. To ad dress this, we establish the first benchmark for noisy labels in Med-VQA by simu lating human mislabeling with semantically designed noise types. More important ly, we introduce the DiN framework, which leverages a diffusion model to handle noisy labels in Med-VQA. Unlike the dominant classification-based VQA approaches that directly predict answers, our Answer Diffuser (AD) module employs a coarse -to-fine process, refining answer candidates with a diffusion model for improved

accuracy. The Answer Condition Generator (ACG) further enhances this process by generating task-specific conditional information via integrating answer embeddings with fused image-question features. To address label noise, our Noisy Label Refinement(NLR) module introduces a robust loss function and dynamic answer adjustment to further boost the performance of the AD module. Our DiN framework consistently outperforms existing methods across multiple benchmarks with varying noise levels.
****
Splatter-360: Generalizable 360 Gaussian Splatting for Wide-baseline Panoramic Images
Zheng Chen, Chenming Wu, Zhelun Shen, Chen Zhao, Weicai Ye, Haocheng Feng, Errui Ding, Song-Hai Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21590-21599
Wide-baseline panoramic images are frequently used in applications like VR and simulations to minimize capturing labor costs and storage needs. However, synthesizing novel views from these panoramic images in real time remains a significant challenge, especially due to panoramic imagery's high resolution and inherent distortions. Although existing 3D Gaussian splatting (3DGS) methods can produce photo-realistic views under narrow baselines, they often overfit the training views when dealing with wide-baseline panoramic images due to the difficulty in learning precise geometry from sparse 360^ \circ views. This paper presents Splatter-360, a novel end-to-end generalizable 3DGS framework designed to handle wide-baseline panoramic images. Unlike previous approaches, Splatter-360 performs multi-view matching directly in the spherical domain by constructing a spherical cost volume through a spherical sweep algorithm, enhancing the network's depth perception and geometry estimation. Additionally, we introduce a 3D-aware bi-projection encoder to mitigate the distortions inherent in panoramic images and integrate cross-view attention to improve feature interactions across multiple viewpoints. This enables robust 3D-aware feature representations and real-time rendering capabilities. Experimental results on the HM3D and Replica demonstrate that Splatter-360 significantly outperforms state-of-the-art NeRF and 3DGS methods (e.g., PanoGRF, MVSplat, DepthSplat, and HiSplat) in both synthesis quality and generalization performance for wide-baseline panoramic images. Code and trained models are available at https://3d-aigc.github.io/Splatter-360/.
****
ShowMak3r: Compositional TV Show Reconstruction
Sangmin Kim, Seunguk Do, Jaesik Park; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 864-874
Reconstructing dynamic radiance fields from video clips is challenging, especially when entertainment videos like TV shows are given. Many challenges make the reconstruction difficult due to (1) actors occluding with each other and having diverse facial expressions, (2) cluttered stages, and (3) small baseline views or sudden shot changes. To address these issues, we present ShowMak3r, a comprehensive reconstruction pipeline that allows the editing of scenes like how video clips are made in a production control room. In ShowMak3r, a 3DLocator module locates recovered actors on the stage using depth prior, and estimates unseen human poses via interpolation. The proposed ShotMatcher module then tracks the actors under shot changes. Furthermore, ShowMak3r introduces a face-fitting network that dynamically recovers the actors' expressions. Experiments on Sitcoms3D dataset show that our pipeline can reassemble TV show scenes with new cameras at different timestamps. We also demonstrate that ShowMak3r enables interesting applications such as synthetic shot-making, actor relocation, insertion, deletion, and pose manipulation. Project page : https://nstar1125.github.io/showmak3r
****
CADRef: Robust Out-of-Distribution Detection via Class-Aware Decoupled Relative Feature Leveraging
Zhiwei Ling, Yachen Chang, Hailiang Zhao, Xinkui Zhao, Kingsum Chow, Shuiguang Deng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4968-4977
Deep neural networks (DNNs) have been widely criticized for their overconfidence

when dealing with out-of-distribution (OOD) samples, highlighting the critical need for effective OOD detection to ensure the safe deployment of DNNs in real-world settings. Existing post-hoc OOD detection methods primarily enhance the discriminative power of logit-based approaches by reshaping sample features, yet they often neglect critical information inherent in the features themselves. In this paper, we propose the \underline C lass-\underline A ware \underline Re lative \underline F eature-based method (CARef), which utilizes the error between a sample's feature and its class-aware average feature as a discriminative criterion. To further refine this approach, we introduce the \underline C lass-\underline A ware \underline D ecoupled \underline Re lative \underline F eature-based method (CADRef), which decouples sample features based on the alignment of signs between the relative feature and corresponding model weights, enhancing the discriminative capabilities of CARef.Extensive experimental results across multiple datasets and models demonstrate that both proposed methods exhibit effectiveness and robustness in OOD detection compared to state-of-the-art methods. Specifically, our two methods outperform the best baseline by 2.82% and 3.27% in AUROC, with improvements of 4.03% and 6.32% in FPR95, respectively.

****

S^3-Face: SSS-Compliant Facial Reflectance Estimation via Diffusion Priors
Xingyu Ren, Jiankang Deng, Yuhao Cheng, Wenhan Zhu, Yichao Yan, Xiaokang Yang, Stefanos Zafeiriou, Chao Ma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16051-16060

Recent 3D face reconstruction methods have made remarkable advancements, yet achieving high-quality facial reflectance from monocular input remains challenging. Existing methods rely on the light-stage captured data to learn facial reflectance models. However, limited subject diversity in these datasets poses challenges in achieving good generalization and broad applicability. This motivates us to explore whether the extensive priors captured in recent generative diffusion models (e.g., Stable Diffusion) can enable more generalizable facial reflectance estimation as these models have been pre-trained on large-scale internet image collections containing rich visual patterns. In this paper, we introduce the use of Stable Diffusion as a prior for facial reflectance estimation, achieving robust results with minimal captured data for fine-tuning. We present S^3-Face, a comprehensive framework capable of producing SSS-compliant skin reflectance from in-the-wild images. Our method adopts a two-stage training approach: in the first stage, DSN-Net is trained to predict diffuse albedo, specular albedo, and normal maps from in-the-wild images using a novel joint reflectance attention module. In the second stage, HM-Net is trained to generate hemoglobin and melanin maps based on the diffuse albedo predicted in the first stage, yielding SSS-compliant and detailed reflectance maps. Extensive experiments demonstrate that our method achieves strong generalization and produces high-fidelity, SSS-compliant facial reflectance reconstructions.

****

FSBench: A Figure Skating Benchmark for Advancing Artistic Sports Understanding
Rong Gao, Xin Liu, Zhuozhao Hu, Bohao Xing, Baiqiang Xia, Zitong Yu, Heikki Kälviäinen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13595-13605

Figure skating, known as the "Art on Ice," is among the most artistic sports, challenging to understand due to its blend of technical elements (like jumps and spins) and overall artistic expression. Existing figure skating datasets mainly focus on single tasks, such as action recognition or scoring, lacking comprehensive annotations for both technical and artistic evaluation. Current sports research is largely centered on ball games, with limited relevance to artistic sports like figure skating. To address this, we introduce FSAnno, a large-scale dataset advancing artistic sports understanding through figure skating. FSAnno includes an open-access training and test dataset, alongside a benchmark dataset, FSBench, for fair model evaluation. FSBench consists of FSBench-Text, with multiple-choice questions and explanations, and FSBench-Motion, containing multimodal data and Question and Answer (QA) pairs, supporting tasks from technical analysis to performance commentary. Initial tests on FSBench reveal significant limitations

in existing models' understanding of artistic sports. We hope FSBench will become a key tool for evaluating and enhancing model comprehension of figure skating. All data, models, and more details are available at: https://github.com/Moomin-Fin/Ano.
****
Keep the Balance: A Parameter-Efficient Symmetrical Framework for RGB+X Semantic Segmentation
Jiaxin Cai, Jingze Su, Qi Li, Wenjie Yang, Shu Wang, Tiesong Zhao, Shengfeng He, Wenxi Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10587-10598
Multimodal semantic segmentation is a critical challenge in computer vision, with early methods suffering from high computational costs and limited transferability due to full fine-tuning of RGB-based pre-trained parameters. Recent studies, while leveraging additional modalities as supplementary prompts to RGB, still predominantly rely on RGB, which restricts the full potential of other modalities. To address these issues, we propose a novel symmetric parameter-efficient fine-tuning framework for multimodal segmentation, featuring with a modality-aware prompting and adaptation scheme, to simultaneously adapt the capabilities of a powerful pre-trained model to both RGB and X modalities. Furthermore, prevalent approaches use the global cross-modality correlations of attention mechanism for modality fusion, which inadvertently introduces noise across modalities. To mitigate this noise, we propose a dynamic sparse cross-modality fusion module to facilitate effective and efficient cross-modality fusion. To further strengthen the above two modules, we propose a training strategy that leverages accurately predicted dual-modality results to self-teach the single-modality outcomes. In comprehensive experiments, we demonstrate that our method outperforms previous state-of-the-art approaches across six multimodal segmentation scenarios with minimal computation cost.
****
VideoDirector: Precise Video Editing via Text-to-Video Models
Yukun Wang, Longguang Wang, Zhiyuan Ma, Qibin Hu, Kai Xu, Yulan Guo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2589-2598
Despite the typical inversion-then-editing paradigm using text-to-image (T2I) models has demonstrated promising results, directly extending it to text-to-video (T2V) models still suffers severe artifacts such as color flickering and content distortion. Consequently, current video editing methods primarily rely on T2I models, which inherently lack temporal-coherence generative ability, often resulting in inferior editing results. In this paper, we attribute the failure of the typical editing paradigm to: 1) Tightly Spatial-temporal Coupling. The vanilla pivotal-based inversion strategy struggles to disentangle spatial-temporal information in the video diffusion model; 2) Complicated Spatial-temporal Layout. The vanilla cross-attention control strategy is deficient in preserving the unedited content. To address these limitations, we propose a spatial-temporal decoupled guidance (STDG) and multi-frame null-text optimization strategy to provide pivotal temporal cues for more precise pivotal inversion. Furthermore, we introduce a self-attention control strategy to maintain higher fidelity for precise partial content editing. Experimental results demonstrate that our method (termed VideoDirector) effectively harnesses the powerful temporal generation capabilities of T2V models, producing edited videos with state-of-the-art performance in accuracy, motion smoothness, realism, and fidelity to unedited content.
****
LLM-driven Multimodal and Multi-Identity Listening Head Generation
Peiwen Lai, Weizhi Zhong, Yipeng Qin, Xiaohang Ren, Baoyuan Wang, Guanbin Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10656-10666
Generating natural listener responses in conversational scenarios is crucial for creating engaging digital humans and avatars. Recent work has shown that large language models (LLMs) can be effectively leveraged for this task, demonstrating remarkable capabilities in generating contextually appropriate listener behavio

rs. However, current LLM-based methods face two critical limitations: they rely solely on speech content, overlooking other crucial communication signals, and they entangle listener identity with response generation, compromising output fidelity and generalization. In this work, we present a novel framework that addresses these limitations while maintaining the advantages of LLMs. Our approach introduces a Multimodal-LM architecture that jointly processes speech content, acoustics, and speaker emotion, capturing the full spectrum of communication cues. Additionally, we propose an identity disentanglement strategy using instance normalization and adaptive instance normalization in a VQ-VAE framework, enabling high-fidelity listening head synthesis with flexible identity control. Extensive experiments demonstrate that our method significantly outperforms existing approaches in terms of response naturalness and fidelity, while enabling effective identity control without retraining.

****

Towards Understanding How Knowledge Evolves in Large Vision-Language Models

Sudong Wang, Yunjian Zhang, Yao Zhu, Jianing Li, Zizhe Wang, Yanwei Liu, Xiangyang Ji; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29858-29868

Large Vision-Language Models (LVLMs) are gradually becoming the foundation for many artificial intelligence applications. However, understanding their internal working mechanisms has continued to puzzle researchers, which in turn limits the further enhancement of their capabilities. In this paper, we seek to investigate how multimodal knowledge evolves and eventually induces natural languages in LVLMs. We design a series of novel strategies for analyzing internal knowledge within LVLMs, and delve into the evolution of multimodal knowledge from three levels, including single token probabilities, token probability distributions, and feature encodings. In this process, we identify two key nodes in knowledge evolution: the critical layers and the mutation layers, dividing the evolution process into three stages: rapid evolution, stabilization, and mutation. Our research is the first to reveal the trajectory of knowledge evolution in LVLMs, providing a fresh perspective for understanding their underlying mechanisms.

****

A Unified, Resilient, and Explainable Adversarial Patch Detector

Vishesh Kumar, Akshay Agarwal; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30387-30397

Deep Neural Networks (DNNs), backbone architecture in `almost' every computer vision task, are vulnerable to adversarial attacks, particularly physical out-of-distribution (OOD) adversarial patches. Existing defense models often struggle with interpreting these attacks in ways that align with human visual perception. Our proposed AdvPatchXAI approach introduces a generalized, robust, and explainable defense algorithm designed to defend DNNs against physical adversarial threats. AdvPatchXAI employs a novel patch decorrelation loss that reduces feature redundancy and enhances the distinctiveness of patch representations, enabling better generalization across unseen adversarial scenarios. It learns prototypical parts self-supervised, enhancing interpretability and correlation with human vision. The model utilizes a sparse linear layer for classification, making the decision process globally interpretable through a set of learned prototypes and locally explainable by pinpointing relevant prototypes within an image. Our comprehensive evaluation shows that AdvPatchXAI closes the "semantic" gap between latent space and pixel space and effectively handles unseen adversarial patches even perturbed with unseen corruptions, thereby significantly advancing DNN robustness in practical settings(https://github.com/tbvl22/Unified-resilient-and-Explainable-Adversarial-Patch-detector).

****

VISTA: Enhancing Long-Duration and High-Resolution Video Understanding by Video Spatiotemporal Augmentation

Weiming Ren, Huan Yang, Jie Min, Cong Wei, Wenhu Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3804-3814

Current large multimodal models (LMMs) face significant challenges in processing and comprehending long-duration or high-resolution videos, which is mainly due

to the lack of high-quality datasets. To address this issue from a data-centric perspective, we propose VISTA, a simple yet effective video spatiotemporal augmentation framework that synthesizes long-duration and high-resolution video instruction-following pairs from existing video-caption datasets. VISTA spatially and temporally combines videos to create new synthetic videos with extended durations and enhanced resolutions, and subsequently produces question-answer pairs pertaining to these newly synthesized videos. Based on this paradigm, we develop seven video augmentation methods and curate VISTA-400K, a video instruction-following dataset aimed at enhancing long-duration and high-resolution video understanding. Finetuning various video LMMs on our data resulted in an average improvement of 3.3% across four challenging benchmarks for long-video understanding. Furthermore, we introduce the first comprehensive high-resolution video understanding benchmark HRVideoBench, on which our finetuned models achieve a 6.5% performance gain. These results highlight the effectiveness of our framework.

****

Structured 3D Latents for Scalable and Versatile 3D Generation

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, Jiaolong Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21469-21480

We introduce a novel 3D generation method for versatile and high-quality 3D asset creation.The cornerstone is a unified Structured LATent (SLAT) representation which allows decoding to different output formats, such as Radiance Fields, 3D Gaussians, and meshes. This is achieved by integrating a sparsely-populated 3D grid with dense multiview visual features extracted from a powerful vision foundation model, comprehensively capturing both structural (geometry) and textural (appearance) information while maintaining flexibility during decoding.We employ rectified flow transformers tailored for SLAT as our 3D generation models and train models with up to 2 billion parameters on a large 3D asset dataset of 500K diverse objects. Our model generates high-quality results with text or image conditions, significantly surpassing existing methods, including recent ones at similar scales. We showcase flexible output format selection and local 3D editing capabilities which were not offered by previous models. Code, model, and data will be released.

****

GA3CE: Unconstrained 3D Gaze Estimation with Gaze-Aware 3D Context Encoding

Yuki Kawana, Shintaro Shiba, Quan Kong, Norimasa Kobori; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3081-3090

We propose a novel 3D gaze estimation approach that learns spatial relationships between the subject and objects in the scene, and outputs 3D gaze direction. Our method targets unconstrained settings, including cases where close-up views of the subject's eyes are unavailable, such as when the subject is distant or facing away. Previous approaches typically rely on either 2D appearance alone or incorporate limited spatial cues using depth maps in the non-learnable post-processing step. Estimating 3D gaze direction from 2D observations in these scenarios is challenging; variations in subject pose, scene layout, and gaze direction, combined with differing camera poses, yield diverse 2D appearances and 3D gaze directions even when targeting the same 3D scene. To address this issue, we propose GA3CE: Gaze-Aware 3D Context Encoding. Our method represents subject and scene using 3D poses and object positions, treating them as 3D context to learn spatial relationships in 3D space. Inspired by human vision, we align this context in an egocentric space, significantly reducing spatial complexity. Furthermore, we propose D^3 (direction-distance-decomposed) positional encoding to better capture the spatial relationship between 3D context and gaze direction in direction and distance space. Experiments demonstrate substantial improvements, reducing mean angle error by 13%-37% compared to leading baselines on benchmark datasets in single-frame settings.

****

Self-Cross Diffusion Guidance for Text-to-Image Synthesis of Similar Subjects

Weimin Qiu, Jieke Wang, Meng Tang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23528-23538

Diffusion models achieved unprecedented fidelity and diversity for synthesizing image, video, 3D assets, etc. However, subject mixing is an unresolved issue for diffusion-based image synthesis, particularly for synthesizing multiple similar-looking subjects. We propose Self-Cross Diffusion Guidance to penalize the overlap between cross-attention maps and the aggregated self-attention map. Compared to previous methods based on self-attention or cross-attention alone, our guidance is more effective in eliminating subject mixing. What's more, our guidance addresses subject mixing for all relevant patches beyond the most discriminant one, e.g., the beak of a bird. For each subject, we aggregate self-attention maps of patches with higher cross-attention values. Thus, the aggregated self-attention map forms a region that the whole subject attends to. Our training-free method boosts the performance of both Unet-based and Transformer-based diffusion models such as the Stable Diffusion series. We also release a similar subjects dataset (SSD), a challenging benchmark, and utilize GPT-4o for automatic and reliable evaluation. Extensive qualitative and quantitative results demonstrate the effectiveness of our self-cross diffusion guidance.

****

RigGS: Rigging of 3D Gaussians for Modeling Articulated Objects in Videos

Yuxin Yao, Zhi Deng, Junhui Hou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5592-5601

This paper considers the problem of modeling articulated objects captured in 2D videos to enable novel view synthesis, while also being easily editable, drivable, and reposable. To tackle this challenging problem, we propose RigGS, a new paradigm that leverages 3D Gaussian representation and skeleton-based motion representation to model dynamic objects without utilizing additional template priors. Specifically, we first propose skeleton-aware node-controlled deformation, which deforms a canonical 3D Gaussian representation over time to initialize the modeling process, producing candidate skeleton nodes that are further simplified into a sparse 3D skeleton according to their motion and semantic information. Subsequently, based on the resulting skeleton, we design learnable skin deformations and pose-dependent detailed deformations, thereby easily deforming the 3D Gaussian representation to generate new actions and render further high-quality images from novel views. Extensive experiments demonstrate that our method can generate realistic new actions easily for objects and achieve high-quality rendering.

****

Noise Modeling in One Hour: Minimizing Preparation Efforts for Self-supervised Low-Light RAW Image Denoising

Feiran Li, Haiyang Jiang, Daisuke Iso; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5699-5708

Noise synthesis is a promising solution for addressing the data shortage problem in data-driven low-light RAW image denoising. However, accurate noise synthesis methods often necessitate labor-intensive calibration and profiling procedures during preparation, preventing them from landing to practice at scale. This work introduces a practically simple noise synthesis pipeline based on detailed analyses of noise properties and extensive justification of widespread techniques. Compared to other approaches, our proposed pipeline eliminates the cumbersome system gain calibration and signal-independent noise profiling steps, reducing the preparation time for noise synthesis from days to hours. Meanwhile, our method exhibits strong denoising performance, showing an up to 0.54dB PSNR improvement over the current state-of-the-art noise synthesis technique.

****

Adv-CPG: A Customized Portrait Generation Framework with Facial Adversarial Attacks

Junying Wang, Hongyuan Zhang, Yuan Yuan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21001-21010

Recent Customized Portrait Generation (CPG) methods, taking a facial image and a textual prompt as inputs, have attracted substantial attention. Although these methods generate high-fidelity portraits, they fail to prevent the generated portraits from being tracked and misused by malicious face recognition systems. To address this, this paper proposes a Customized Portrait Generation framework wit

h facial Adversarial attacks (Adv-CPG). Specifically, to achieve facial privacy protection, we devise a lightweight local ID encryptor and an encryption enhancer. They implement progressive double-layer encryption protection by directly injecting the target identity and adding additional identity guidance, respectively. Furthermore, to accomplish fine-grained and personalized portrait generation, we develop a multi-modal image customizer capable of generating controlled fine-grained facial features. To the best of our knowledge, Adv-CPG is the first study that introduces facial adversarial attacks into CPG. Extensive experiments demonstrate the superiority of Adv-CPG, e.g., the average attack success rate of the proposed Adv-CPG is 28.1% and 2.86% higher compared to the SOTA noise-based attack methods and unconstrained attack methods, respectively.

****

Fish-Vista: A Multi-Purpose Dataset for Understanding & Identification of Traits from Images

Kazi Sajeed Mehrab, M. Maruf, Arka Daw, Abhilash Neog, Harish Babu Manogaran, Mridul Khurana, Zhenyang Feng, Bahadir Altintas, Yasin Bakis, Elizabeth G Campolongo, Matthew J Thompson, Xiaojun Wang, Hilmar Lapp, Tanya Berger-Wolf, Paula Mabee, Henry Bart, Wei-Lun Chao, Wasila M Dahdul, Anuj Karpatne; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24275-24285

We introduce Fish-Visual Trait Analysis (Fish-Vista), the first organismal image dataset designed for the analysis of visual traits of aquatic species directly from images using machine learning and computer vision methods. Fish-Vista contains 69,269 annotated images spanning 4,316 fish species, curated and organized to serve three downstream tasks: species classification, trait identification, and trait segmentation. Our work makes two key contributions. First, we provide a fully reproducible data processing pipeline to process fish images sourced from various museum collections, contributing to the advancement of AI in biodiversity science. We annotate the images with carefully curated labels from biological databases and manual annotations to create an AI-ready dataset of visual traits. Second, our work offers fertile grounds for researchers to develop novel methods for a variety of problems in computer vision such as handling long-tailed distributions, out-of-distribution generalization, learning with weak labels, explainable AI, and segmenting small objects. Dataset and code for Fish-Vista are available at https://github.com/Imageomics/Fish-Vista

****

High Dynamic Range Video Compression: A Large-Scale Benchmark Dataset and A Learned Bit-depth Scalable Compression Algorithm

Zhaoyi Tian, Feifeng Wang, Shiwei Wang, Zihao Zhou, Yao Zhu, Liquan Shen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7320-7330

Recently, learned video compression (LVC) is undergoing a period of rapid development. However, due to absence of large and high-quality high dynamic range (HDR) video training data, LVC on HDR video is still unexplored. In this paper, we are the first to collect a large-scale HDR video benchmark dataset, named HDRVD2K, featuring huge quantity, diverse scenes and multiple motion types. HDRVD2K fills gaps of video training data and facilitate the development of LVC on HDR videos. Based on HDRVD2K, we further propose the first learned bit-depth scalable video compression (LBSVC) network for HDR videos by effectively exploiting bit-depth redundancy between videos of multiple dynamic ranges. To achieve this, we first propose a compression-friendly bit-depth enhancement module (BEM) to effectively predict original HDR videos based on compressed tone-mapped low dynamic range (LDR) videos and dynamic range prior, instead of reducing redundancy only through spatio-temporal predictions. Our method greatly improves the reconstruction quality and compression performance on HDR videos. Extensive experiments demonstrate the effectiveness of HDRVD2K on learned HDR video compression and great compression performance of our proposed LBSVC network.

****

OffsetOPT: Explicit Surface Reconstruction without Normals

Huan Lei; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11729-11738

Neural surface reconstruction has been dominated by implicit representations wit
h marching cubes for explicit surface extraction. However, those methods typical
ly require high-quality normals for accurate reconstruction. We propose OffsetOP
T, a method that reconstructs explicit surfaces directly from 3D point clouds
and eliminates the need for point normals. The approach comprises two stages: fi
rst, we train a neural network to predict surface triangles based on local point
geometry, given uniformly distributed training point clouds. Next, we apply the
frozen network to reconstruct surfaces from unseen point clouds by optimizing a
per-point offset to maximize the accuracy of triangle predictions. Compared to
state-of-the-art methods, OffsetOPT not only excels at reconstructing overall su
rfaces but also significantly preserves sharp surface features. We demonstrate i
ts accuracy on popular benchmarks, including small-scale shapes and large-scale
open surfaces.
****

PCM : Picard Consistency Model for Fast Parallel Sampling of Diffusion Models
Junhyuk So, Jiwoong Shin, Chaeyeon Jang, Eunhyeok Park; Proceedings of the Compu
ter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23313-23322
Recently, diffusion models have achieved significant advances in vision, text, a
nd robotics. However, they still face slow generation speeds due to sequential d
enoising processes. To address this, a parallel sampling method based on Picard
iteration was introduced, effectively reducing sequential steps while ensuring e
xact convergence to the original output. Nonetheless, Picard iteration does not
guarantee faster convergence, which can still result in slow generation in pract
ice. In this work, we propose a new parallelization scheme, the Picard Consisten
cy Model (PCM), which significantly reduces the number of generation steps in Pi
card iteration. Inspired by the consistency model, PCM is directly trained to pr
edict the fixed-point solution, or the final output, at any stage of the converg
ence trajectory. Additionally, we introduce a new concept called model switching
, which addresses PCM's limitations and ensures exact convergence. Extensive exp
eriments demonstrate that PCM achieves up to a 2.71x speedup over sequential sam
pling and a 1.77x speedup over Picard iteration across various tasks, including
image generation and robotic control.
****

CoMapGS: Covisibility Map-based Gaussian Splatting for Sparse Novel View Synthes
is
Youngkyoon Jang, Eduardo Pérez-Pellitero; Proceedings of the Computer Vision and
Pattern Recognition Conference (CVPR), 2025, pp. 26779-26788
We propose Covisibility Map-based Gaussian Splatting (CoMapGS), designed to reco
ver underrepresented sparse regions in sparse novel view synthesis. CoMapGS addr
esses both high- and low-uncertainty regions by constructing covisibility maps,
enhancing initial point clouds, and applying uncertainty-aware weighted supervis
ion using a proximity classifier. Our contributions are threefold: (1) CoMapGS r
eframes novel view synthesis by leveraging covisibility maps as a core component
to address region-specific uncertainty; (2) Enhanced initial point clouds for b
oth low- and high-uncertainty regions compensate for sparse COLMAP-derived point
clouds, improving reconstruction quality and benefiting few-shot 3DGS methods;
(3) Adaptive supervision with covisibility-score-based weighting and proximity c
lassification achieves consistent performance gains across scenes with varying s
parsity scores derived from covisibility maps. Experimental results demonstrate
that CoMapGS outperforms state-of-the-art methods on datasets including Mip-NeRF
360 and LLFF.
****

Any-Resolution AI-Generated Image Detection by Spectral Learning
Dimitrios Karageorgiou, Symeon Papadopoulos, Ioannis Kompatsiaris, Efstratios Ga
vves; Proceedings of the Computer Vision and Pattern Recognition Conference (CVP
R), 2025, pp. 18706-18717
Recent works have established that AI models introduce spectral artifacts into g
enerated images and propose approaches for learning to capture them using labele
d data. However, the significant differences in such artifacts among different g
enerative models hinder these approaches from generalizing to generators not see

n during training. In this work, we build upon the key idea that the spectral di
stribution of real images constitutes both an invariant and highly discriminativ
e pattern for AI-generated image detection. To model this under a self-supervise
d setup, we employ masked spectral learning using the pretext task of frequency
reconstruction. Since generated images constitute out-of-distribution samples fo
r this model, we propose spectral reconstruction similarity to capture this dive
rgence. Moreover, we introduce spectral context attention, which enables our app
roach to efficiently capture subtle spectral inconsistencies in images of any re
solution. Our spectral AI-generated image detection approach (SPAI) achieves a 5
.5% absolute improvement in AUC over the previous state-of-the-art across 13 rec
ent generative approaches, while exhibiting robustness against common online per
turbations. Code is available on https://mever-team.github.io/spai.
****

DivPrune: Diversity-based Visual Token Pruning for Large Multimodal Models
Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, Yong Zhang; Proceedings o
f the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9392-
9401
Large Multimodal Models (LMMs) have emerged as powerful models capable of unders
tanding various data modalities, including text, images, and videos. LMMs encode
 both text and visual data into tokens that are then combined and processed by a
n integrated Large Language Model (LLM). Including visual tokens substantially i
ncreases the total token count, often by thousands. The increased input length f
or LLM significantly raises the complexity of inference, resulting in high laten
cy in LMMs. To address this issue, token pruning methods, which remove part of t
he visual tokens, are proposed. The existing token pruning methods either requir
e extensive calibration and fine-tuning or rely on suboptimal importance metrics
 which results in increased redundancy among the retained tokens. In this paper,
 we first formulate token pruning as Max-Min Diversity Problem (MMDP) where the
goal is to select a subset such that the diversity among the selected tokens is
maximized. Then, we solve the MMDP to obtain the selected subset and prune the r
est. The proposed method, DivPrune, reduces redundancy and achieves the highest
diversity of the selected tokens. By ensuring high diversity, the selected token
s better represent the original tokens, enabling effective performance even at h
igh pruning ratios without requiring fine-tuning. Extensive experiments with var
ious LMMs show that DivPrune achieves state-of-the-art accuracy over 16 image- a
nd video-language datasets. Additionally, DivPrune reduces both the end-to-end l
atency and GPU memory usage for the tested models. The code is available here.
****

Training Data Provenance Verification: Did Your Model Use Synthetic Data from My
 Generative Model for Training?
Yuechen Xie, Jie Song, Huiqiong Wang, Mingli Song; Proceedings of the Computer V
ision and Pattern Recognition Conference (CVPR), 2025, pp. 23817-23827
High-quality open-source text-to-image models have lowered the threshold for obt
aining photorealistic images significantly, but also face potential risks of mis
use. Specifically, suspects may use synthetic data generated by these generative
 models to train models for specific tasks without permission, when lacking real
 data resources especially. Protecting these generative models is crucial for th
e well-being of their owners. In this work, we propose the first method to this
important yet unresolved issue, called Training data Provenance Verification (Tr
ainProVe). The rationale behind TrainProVe is grounded in the principle of gener
alization error bound, which suggests that, for two models with the same task, i
f the distance between their training data distributions is smaller, their gener
alization ability will be closer. We validate the efficacy of TrainProVe across
four text-to-image models (Stable Diffusion v1.4, latent consistency model, PixA
rt-\alpha, and Stable Cascade). The results show that TrainProVe achieves a veri
fication accuracy of over 99% in determining the provenance of suspicious model
training data, surpassing all previous methods. Code is available at https://git
hub.com/xieyc99/TrainProVe.
****

3D-AVS: LiDAR-based 3D Auto-Vocabulary Segmentation

Weijie Wei, Osman Ülger, Fatemeh Karimi Nejadasl, Theo Gevers, Martin R. Oswald; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8910-8920

Open-vocabulary segmentation methods offer promising capabilities in detecting unseen object categories, but the category must be aware and needs to be provided by a human, either via a text prompt or pre-labeled datasets, thus limiting their scalability. We propose 3D-AVS, a method for Auto-Vocabulary Segmentation of 3D point clouds for which the vocabulary is unknown and auto-generated for each input at runtime, thus eliminating the human in the loop and typically providing a substantially larger vocabulary for richer annotations. 3D-AVS first recognizes semantic entities from image or point cloud data and then segments all points with the automatically generated vocabulary. Our method incorporates both image-based and point-based recognition, enhancing robustness under challenging lighting conditions where geometric information from LiDAR is especially valuable. Our point-based recognition features a Sparse Masked Attention Pooling (SMAP) module to enrich the diversity of recognized objects. To address the challenges of evaluating unknown vocabularies and avoid annotation biases from label synonyms, hierarchies, or semantic overlaps, we introduce the annotation-free Text-Point Semantic Similarity (TPSS) metric for assessing generated vocabulary quality. Our evaluations on nuScenes and ScanNet200 demonstrate 3D-AVS's ability to generate semantic classes with accurate point-wise segmentations.

****

STOP: Integrated Spatial-Temporal Dynamic Prompting for Video Understanding

Zichen Liu, Kunlun Xu, Bing Su, Xu Zou, Yuxin Peng, Jiahuan Zhou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13776-13786

Pre-trained on tremendous image-text pairs, vision-language models like CLIP have demonstrated promising zero-shot generalization across numerous image-based tasks. However, extending these capabilities to video tasks remains challenging due to limited labeled video data and high training costs. Recent video prompting methods attempt to adapt CLIP for video tasks by introducing learnable prompts, but they typically rely on a single static prompt for all video sequences, overlooking the diverse temporal dynamics and spatial variations that exist across frames. This limitation significantly hinders the model's ability to capture essential temporal information for effective video understanding. To address this, we propose an integrated Spatial-TempOral dynamic Prompting (STOP) model which consists of two complementary modules, the intra-frame spatial prompting and inter-frame temporal prompting. Our intra-frame spatial prompts are designed to adaptively highlight discriminative regions within each frame by leveraging intra-frame attention and temporal variation, allowing the model to focus on areas with substantial temporal dynamics and capture fine-grained spatial details. Additionally, to highlight the varying importance of frames for video understanding, we further introduce inter-frame temporal prompts, dynamically inserting prompts between frames with high temporal variance as measured by frame similarity. This enables the model to prioritize key frames and enhances its capacity to understand temporal dependencies across sequences. Extensive experiments on various video benchmarks demonstrate that STOP consistently achieves superior performance against state-of-the-art methods. The code is available at https://github.com/zhoujiahuan1991/CVPR2025-STOP

****

TimeTracker: Event-based Continuous Point Tracking for Video Frame Interpolation with Non-linear Motion

Haoyue Liu, Jinghan Xu, Yi Chang, Hanyu Zhou, Haozhi Zhao, Lin Wang, Luxin Yan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17649-17659

Video frame interpolation (VFI) that leverages the bio-inspired event cameras as guidance has recently shown better performance and memory efficiency than the frame-based methods, thanks to the event cameras' advantages, such as high temporal resolution. A hurdle for event-based VFI is how to effectively deal with non-linear motion, caused by the dynamic changes in motion direction and speed withi

n the scene. Existing methods either use events to estimate sparse optical flow or fuse events with image features to estimate dense optical flow. Unfortunately, motion errors often degrade the VFI quality as the continuous motion cues from events do not align with the dense spatial information of images in the temporal dimension. In this paper, we find that object motion is continuous in space, tracking local regions over continuous time enables more accurate identification of spatiotemporal feature correlations. In light of this, we propose a novel continuous point tracking-based VFI framework, named TimeTracker. Specifically, we first design a Scene-Aware Region Segmentation (SARS) module to divide the scene into similar patches. Then, a Continuous Trajectory guided Motion Estimation (CTME) module is proposed to track the continuous motion trajectory of each patch through events. Finally, intermediate frames at any given time are generated through global motion optimization and frame refinement. Moreover, we collect a real-world dataset that features fast non-linear motion. Extensive experiments show that our method outperforms prior arts in both motion estimation and frame interpolation quality.

****

Improving the Training of Data-Efficient GANs via Quality Aware Dynamic Discriminator Rejection Sampling

Zhaoyu Zhang, Yang Hua, Guanxiong Sun, Hui Wang, Seán McLoone; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30682-30691

Data-Efficient Generative Adversarial Nets (DE-GANs) have become more and more popular in recent years. Existing methods apply data augmentation, noise injection and pre-trained models to maximumly increase the number of training samples thus improving the training of DE-GANs. However, none of these methods considers the sample quality during training, which can also significantly influence the training of DE-GANs. Focusing on sample quality during training, in this paper, we are the first to incorporate discriminator rejection sampling (DRS) into the training process and introduce a novel method, called quality aware dynamic discriminator rejection sampling (QADDRS). Specifically, QADDRS consists of two steps: (1) the sample quality aware step, which aims to obtain the sorted critic scores, i.e., the ordered discriminator outputs, on real/fake samples in the current training stage; (2) the dynamic rejection step that obtains dynamic rejection number N, where N is controlled by the overfitting degree of discriminator (D) during training. When updating the parameters of D, the N high critic score real samples and the N low critic score fake samples in the minibatch are rejected dynamically based on the overfitting degree of D. As a result, QADDRS can avoid D becoming overly confident in distinguishing both real and fake samples, thereby alleviating the overfitting of D issue during training. Extensive experiments on several datasets demonstrate that integrating QADDRS into different DE-GANs can achieve better performance and deliver state-of-the-art results. Codes are available at https://github.com/zzhang05/QADDRS.

****

Shading Meets Motion: Self-supervised Indoor 3D Reconstruction Via Simultaneous Shape-from-Shading and Structure-from-Motion

Guoyu Lu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16508-16519

Scene reconstruction has a wide range of applications in computer vision and robotics. To build practical constraints and feature Scene reconstruction has a wide range of applications in computer vision and robotics. To build practical constraints and feature correspondences, rich textures and distinguished gradient variations are particularly required in classic and learning-based SfM. When building low-texture regions with repeated patterns, especially mostly-white indoor rooms, there is a significant drop in performance. In this work, we propose Shading-SfM-Net, a novel framework for simultaneously learning a shape-from-shading network based on the inverse rendering constraint and a structure-from-motion framework based on warped keypoint and geometric consistency, to improve structure-from-motion and surface reconstruction for low-texture indoor scenes. Shading-SfM-Net tightly incorporates the surface shape consistency and 3D geometric regist

ration loss in order to dig into their mutual information and further overcome t
he instability on flat regions. We evaluate the proposed framework on texture-le
ss indoor scenes (NYUv2 and ScanNet), and show that by simultaneously learning s
hading, motion and shape, our pipeline is able to achieve state-of-the-art perfo
rmance with superior generalization capability for unseen texture-less datasets.
****

Believing is Seeing: Unobserved Object Detection using Generative Models
Subhransu S. Bhattacharjee, Dylan Campbell, Rahul Shome; Proceedings of the Comp
uter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19366-19377
Can objects that are not visible in an image---but are in the vicinity of the ca
mera---be detected? This study introduces the novel tasks of 2D, 2.5D and 3D uno
bserved object detection for predicting the location of nearby objects that are
occluded or lie outside the image frame. We adapt several state-of-the-art pre-
trained generative models to address this task, including 2D and 3D diffusion mo
dels and vision-language models, and show that they can be used to infer the pre
sence of objects that are not directly observed. To benchmark this task, we pro
pose a suite of metrics that capture different aspects of performance. Our empi
rical evaluation on indoor scenes from the RealEstate10k and NYU Depth v2 datase
ts demonstrate results that motivate the use of generative models for the unobse
rved object detection task.
****

MotionStone: Decoupled Motion Intensity Modulation with Diffusion Transformer fo
r Image-to-Video Generation
Shuwei Shi, Biao Gong, Xi Chen, Dandan Zheng, Shuai Tan, Zizheng Yang, Yuyuan Li
, Jingwen He, Kecheng Zheng, Jingdong Chen, Ming Yang, Yinqiang Zheng; Proceedin
gs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2
2864-22874
The image-to-video (I2V) generation is conditioned on the static image, which ha
s been enhanced recently by the motion intensity as an additional control signal
. These motion-aware models are appealing to generate diverse motion patterns, y
et there lacks a reliable motion estimator for training such models on large-sca
le video set in the wild. Traditional metrics, e.g., SSIM or optical flow, are h
ard to generalize to arbitrary videos, while, it is very tough for human annotat
ors to label the abstract motion intensity neither. Furthermore, the motion inte
nsity shall reveal both local object motion and global camera movement, which ha
s not been studied before. This paper addresses the challenge with a new motion
estimator, capable of measuring the decoupled motion intensities of objects and
cameras in video. We leverage the contrastive learning on randomly paired videos
 and distinguish the video with greater motion intensity. Such a paradigm is fri
endly for annotation and easy to scale up to achieve stable performance on motio
n estimation. We then present a new I2V model, named MotionStone, developed with
 the decoupled motion estimator. Experimental results demonstrate the stability
of the proposed motion estimator and the state-of-the-art performance of MotionS
tone on I2V generation. These advantages warrant the decoupled motion estimator
to serve as a general plug-in enhancer for both data processing and video genera
tion training.
****

NLPrompt: Noise-Label Prompt Learning for Vision-Language Models
Bikang Pan, Qun Li, Xiaoying Tang, Wei Huang, Zhen Fang, Feng Liu, Jingya Wang,
Jingyi Yu, Ye Shi; Proceedings of the Computer Vision and Pattern Recognition Co
nference (CVPR), 2025, pp. 19963-19973
The emergence of vision-language foundation models, such as CLIP, has revolution
ized image-text representation, enabling a broad range of applications via promp
t learning. Despite its promise, real-world datasets often contain noisy labels
that can degrade prompt learning performance. In this paper, we demonstrate that
 using mean absolute error (MAE) loss in prompt learning, named PromptMAE, signi
ficantly enhances robustness against noisy labels while maintaining high accurac
y. Though MAE is straightforward and recognized for its robustness, it is rarely
 used in noisy-label learning due to its slow convergence and poor performance o
utside prompt learning scenarios. To elucidate the robustness of PromptMAE, we l

everage feature learning theory to show that MAE can suppress the influence of n
oisy samples, thereby improving the signal-to-noise ratio and enhancing overall
robustness. Additionally, we introduce PromptOT, a prompt-based optimal transpor
t data purification method to enhance the robustness further. PromptOT employs t
ext encoder representations in vision-language models as prototypes to construct
 an optimal transportation matrix. This matrix effectively partitions datasets i
nto clean and noisy subsets, allowing for the application of cross-entropy loss
to the clean subset and MAE loss to the noisy subset. Our Noise-Label Prompt Lea
rning method, named NLPrompt, offers a simple and efficient approach that levera
ges the expressive representation and precise alignment capabilities of vision-l
anguage models for robust prompt learning. We validate NLPrompt through extensi
ve experiments across various noise settings, demonstrating significant performa
nce improvements.
****

MEGA: Masked Generative Autoencoder for Human Mesh Recovery
Guénolé Fiche, Simon Leglaive, Xavier Alameda-Pineda, Francesc Moreno-Noguer; Pr
oceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025
, pp. 5366-5378
Human Mesh Recovery (HMR) from a single RGB image is a highly ambiguous problem,
 as an infinite set of 3D interpretations can explain the 2D observation equally
 well. Nevertheless, most HMR methods overlook this issue and make a single pred
iction without accounting for this ambiguity. A few approaches generate a distri
bution of human meshes, enabling the sampling of multiple predictions; however,
none of them is competitive with the latest single-output model when making a si
ngle prediction. This work proposes a new approach based on masked generative mo
deling. By tokenizing the human pose and shape, we formulate the HMR task as gen
erating a sequence of discrete tokens conditioned on an input image. We introduc
e MEGA, a MaskEd Generative Autoencoder trained to recover human meshes from ima
ges and partial human mesh token sequences. Given an image, our flexible generat
ion scheme allows us to predict a single human mesh in deterministic mode or to
generate multiple human meshes in stochastic mode. Experiments on in-the-wild be
nchmarks show that MEGA achieves state-of-the-art performance in deterministic a
nd stochastic modes, outperforming single-output and multi-output approaches.
****

PBR-NeRF: Inverse Rendering with Physics-Based Neural Fields
Sean Wu, Shamik Basu, Tim Broedermann, Luc Van Gool, Christos Sakaridis; Proceed
ings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp.
 10974-10984
We tackle the ill-posed inverse rendering problem in 3D reconstruction with a Ne
ural Radiance Field (NeRF) approach informed by Physics-Based Rendering (PBR) th
eory, named PBR-NeRF. Our method addresses a key limitation in most NeRF and 3D
Gaussian Splatting approaches: they estimate view-dependent appearance without m
odeling scene materials and illumination. To address this limitation, we present
 an inverse rendering (IR) model capable of jointly estimating scene geometry, m
aterials, and illumination. Our model builds upon recent NeRF-based IR approache
s, but crucially introduces two novel physics-based priors that better constrain
 the IR estimation. Our priors are rigorously formulated as intuitive loss terms
 and achieve state-of-the-art material estimation without compromising novel vie
w synthesis quality. Our method is easily adaptable to other inverse rendering a
nd 3D reconstruction frameworks that require material estimation. We demonstrate
 the importance of extending current neural rendering approaches to fully model
scene properties beyond geometry and view-dependent appearance. Code is publicly
 available at: https://github.com/s3anwu/pbrnerf.
****

Disentangling Safe and Unsafe Image Corruptions via Anisotropy and Locality
Ramchandran Muthukumar, Ambar Pal, Jeremias Sulam, Rene Vidal; Proceedings of th
e Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9954-9963
State-of-the-art machine learning systems are vulnerable to small perturbations
to their input, where _small_ is defined according to a threat model that assign
s a positive threat to each perturbation. Most prior works define a task-agnosti

c, isotropic, and global threat, like the l_p norm, where the magnitude of the p erturbation fully determines the degree of the threat and neither the direction of the attack nor its position in space matter. However, common corruptions in c omputer vision, such as blur, compression, or occlusions, are not well captured by such treat models. This paper proposes a novel threat model called \texttt Pr ojected Displacement (PD) to study robustness beyond existing isotropic and glo bal threat models. The proposed threat model measures the threat of a perturbati on via its alignment with _unsafe directions_, defined as directions in the inpu t space along which a perturbation of sufficient magnitude changes the ground tr uth class label. Unsafe directions are identified locally for each input based o n observed training data. In this way, the PD-threat model exhibits anisotropy a nd locality. The PD-threat model is computationally efficient and can be easily integrated into existing robustness pipelines. Experiments on Imagenet-1k data i ndicate that, for any input, the set of perturbations with small PD threat inclu des _safe_ perturbations of large l_p norm that preserve the true label, such as noise, blur and compression, while simultaneously excluding _unsafe_ perturbati ons that alter the true label. Unlike perceptual threat models based on embeddin gs of large-vision models, the PD-threat model can be readily computed for arbit rary classification tasks without pre-training or finetuning. Further additional task information such as sensitivity to image regions or concept hierarchies ca n be easily integrated into the assessment of threat and thus the PD threat mode l presents practitioners a flexible, task-driven threat specification that allev iates the limitations of l_p-threat models.

****

Prometheus: 3D-Aware Latent Diffusion Models for Feed-Forward Text-to-3D Scene G eneration

Yuanbo Yang, Jiahao Shao, Xinyang Li, Yujun Shen, Andreas Geiger, Yiyi Liao; Pro ceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2857-2869

In this work, we introduce Prometheus, a 3D-aware latent diffusion model for tex t-to-3D generation at both object and scene levels in seconds. We formulate 3D s cene generation as multi-view, feed-forward, pixel-aligned 3D Gaussian generatio n within the latent diffusion paradigm. To ensure generalizability, we build our model upon pre-trained text-to-image generation model with only minimal adjustm ents and further train it using a large number of images from both single-view a nd multi-view datasets. Furthermore, we introduce an RGB-D latent space into 3D Gaussian generation to disentangle appearance and geometry information, enabling efficient feed-forward generation of 3D Gaussians with better fidelity and geom etry. Extensive experimental results demonstrate the effectiveness of our method in both feed-forward 3D Gaussian reconstruction and text-to-3D generation.

****

No Pains, More Gains: Recycling Sub-Salient Patches for Efficient High-Resolutio n Image Recognition

Rong Qin, Xin Liu, Xingyu Liu, Jiaxuan Liu, Jinglei Shi, Liang Lin, Jufeng Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2 025, pp. 14965-14975

Over the last decade, many notable methods have emerged to tackle the computatio nal resource challenge of the high resolution image recognition (HRIR). They typ ically focus on identifying and aggregating a few salient regions for classifica tion, discarding sub-salient areas for low training consumption. Nevertheless, m any HRIR tasks necessitate the exploration of wider regions to model objects and contexts, which limits their performance in such scenarios. To address this iss ue, we present a DBPS strategy to enable training with more patches at low consu mption. Specifically, in addition to a fundamental buffer that stores the embedd ings of most salient patches, DBPS further employs an auxiliary buffer to recycl e those sub-salient ones. To reduce the computational cost associated with gradi ents of sub-salient patches, these patches are primarily used in the forward pas s to provide sufficient information for classification. Meanwhile, only the grad ients of the salient patches are back-propagated to update the entire network. M oreover, we design a Multiple Instance Learning (MIL) architecture that leverage

s aggregated information from salient patches to filter out uninformative backgr ound within sub-salient patches for better accuracy. Besides, we introduce the r andom patch drop to accelerate training process and uncover informative regions. Experiment results demonstrate the superiority of our method in terms of both a ccuracy and training consumption against other advanced methods. The code is ava ilable in the https://github.com/Qinrong-NKU/DBPS.
****

SphereUFormer: A U-Shaped Transformer for Spherical 360 Perception
Yaniv Benny, Lior Wolf; Proceedings of the Computer Vision and Pattern Recogniti on Conference (CVPR), 2025, pp. 940-950
This paper proposes a novel method for omnidirectional 360\degree perception. Mo st common previous methods relied on equirectangular projection. This representa tion is easily applicable to 2D operation layers but introduces distortions into the image. Other methods attempted to remove the distortions by maintaining a s phere representation but relied on complicated convolution kernels that failed t o show competitive results. In this work, we introduce a transformer-based archi tecture that, by incorporating a novel "Spherical Local Self-Attention" and othe r spherically-oriented modules, successfully operates in the spherical domain an d outperforms the state-of-the-art in 360\degree perception benchmarks for depth estimation and semantic segmentation. Our code is available at https://github.c om/yanivbenny/sphere_uformer.
****

Advancing Generalizable Tumor Segmentation with Anomaly-Aware Open-Vocabulary At tention Maps and Frozen Foundation Diffusion Models
Yankai Jiang, Peng Zhang, Donglin Yang, Yuan Tian, Hai Lin, Xiaosong Wang; Proce edings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, p p. 25971-25981
We explore Generalizable Tumor Segmentation, aiming to train a single model for zero-shot tumor segmentation across diverse anatomical regions. Existing methods face limitations related to segmentation quality, scalability, and the range of applicable imaging modalities. In this paper, we uncover the potential of the i nternal representations within frozen medical foundation diffusion models as hig hly efficient zero-shot learners for tumor segmentation by introducing a novel f ramework named DiffuGTS. DiffuGTS creates anomaly-aware open-vocabulary attentio n maps based on text prompts to enable generalizable anomaly segmentation withou t being restricted by a predefined training category list. To further improve an d refine anomaly segmentation masks, DiffuGTS leverages the diffusion model, tra nsforming pathological regions into high-quality pseudo-healthy counterparts thr ough latent space inpainting, and applies a novel pixel-level and feature-level residual learning approach, resulting in segmentation masks with significantly e nhanced quality and generalization. Comprehensive experiments on four datasets a nd seven tumor categories demonstrate the superior performance of our method, su rpassing current state-of-the-art models across multiple zero-shot settings. Cod es are available at https://github.com/Yankai96/DiffuGTS.
****

Towards Generalizable Scene Change Detection
Jae-Woo Kim, Ue-Hwan Kim; Proceedings of the Computer Vision and Pattern Recogni tion Conference (CVPR), 2025, pp. 24463-24473
While current state-of-the-art Scene Change Detection (SCD) approaches achieve i mpressive results in well-trained research data, they become unreliable under un seen environments and different temporal conditions; in-domain performance drops from 77.6% to 8.0% in a previously unseen environment and to 4.6% under a diffe rent temporal condition---calling for generalizable SCD and benchmark. In this w ork, we propose the Generalizable Scene Change Detection Framework (GeSCF), whic h addresses unseen domain performance and temporal consistency---to meet the gro wing demand for anything SCD. Our method leverages the pre-trained Segment Anyth ing Model (SAM) in a zero-shot manner. For this, we design Initial Pseudo-mask G eneration and Geometric-Semantic Mask Matching---seamlessly turning user-guided prompt and single-image based segmentation into scene change detection for a pai r of inputs without guidance. Furthermore, we define the Generalizable Scene Cha

nge Detection (GeSCD) benchmark along with novel metrics and an evaluation proto col to facilitate SCD research in generalizability. In the process, we introduce the ChangeVPR dataset, a collection of challenging image pairs with diverse env ironmental scenarios---including urban, suburban, and rural settings. Extensive experiments across various datasets demonstrate that GeSCF achieves an average p erformance gain of 19.2% on existing SCD datasets and 30.0% on the ChangeVPR dat aset, nearly doubling the prior art performance. We believe our work can lay a s olid foundation for robust and generalizable SCD research.

****

Beyond Clean Training Data: A Versatile and Model-Agnostic Framework for Out-of-Distribution Detection with Contaminated Training Data

Yuchuan Li, Jae-Mo Kang, Il-Min Kim; Proceedings of the Computer Vision and Patt ern Recognition Conference (CVPR), 2025, pp. 10183-10192

In real-world AI applications, training datasets are often contaminated, contain ing a mix of in-distribution (ID) and out-of-distribution (OOD) samples without labels. This contamination poses a significant challenge for developing and trai ning OOD detection models, as nearly all existing methods assume access to a cle an training dataset of only ID samples--a condition rarely met in real-world sce narios. Customizing each existing OOD detection method to handle such contaminat ion is impractical, given the vast number of diverse methods designed for clean data. To address this issue, we propose a universal, model-agnostic framework th at integrates with nearly all existing OOD detection methods, enabling training on contaminated datasets while achieving high OOD detection accuracy on test dat asets. Additionally, our framework provides an accurate estimation of the unknow n proportion of OOD samples within the training dataset--an important and distin ct challenge in its own right. Our approach introduces a novel dynamic weighting function and transition mechanism within an iterative training structure, enabl ing both reliable estimation of the OOD sample proportion of the training data a nd precise OOD detection on test data. Extensive evaluations across diverse data sets, including ImageNet-1k, demonstrate that our framework accurately estimates OOD sample proportions of training data and substantially enhances OOD detectio n accuracy on test data.

****

Incomplete Multi-modal Brain Tumor Segmentation via Learnable Sorting State Spac e Model

Zheyu Zhang, Yayuan Lu, Feipeng Ma, Yueyi Zhang, Huanjing Yue, Xiaoyan Sun; Proc eedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25982-25992

Brain tumor segmentation plays a crucial role in clinical diagnosis, yet the fre quent unavailability of certain MRI modalities poses a significant challenge. In this paper, we introduce the Learnable Sorting State Space Model (LS3M), a nove l framework designed to maximize the utilization of available modalities for bra in tumor segmentation. LS3M excels at efficiently modeling long-range dependenci es based on the Mamba design, while incorporating differentiable permutation mat rices that reorder input sequences based on modality-specific characteristics. T his dynamic reordering ensures that critical spatial inductive biases and long-r ange semantic correlations inherent in 3D brain MRI are preserved, which is cruc ial for imcomplete multi-modal brain tumor segmentation.Once the input sequences are reordered using the generated permutation matrix, the Series State Space Mo del (S3M) block models the relationships between them, capturing both local and long-range dependencies. This enables effective representation of intra-modal an d inter-modal relationships, significantly improving segmentation accuracy. Exte nsive experiments on the BraTS2018 and BraTS2020 datasets demonstrate that LS3M outperforms existing methods, offering a robust solution for brain tumor segment ation, particularly in scenarios with missing modalities.

****

FedAWA: Adaptive Optimization of Aggregation Weights in Federated Learning Using Client Vectors

Changlong Shi, He Zhao, Bingjie Zhang, Mingyuan Zhou, Dandan Guo, Yi Chang; Proc eedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025,

pp. 30651-30660
Federated Learning (FL) has emerged as a promising framework for distributed machine learning, enabling collaborative model training without sharing local data, thereby preserving privacy and enhancing security. However, data heterogeneity resulting from differences across user behaviors, preferences, and device characteristics poses a significant challenge for federated learning. Most previous works overlook the adjustment of aggregation weights, relying solely on dataset size for weight assignment, which often leads to unstable convergence and reduced model performance. Recently, several studies have sought to refine aggregation strategies by incorporating dataset characteristics and model alignment. However, adaptively adjusting aggregation weights while ensuring data security--without requiring additional proxy data--remains a significant challenge. In this work, we propose Federated learning with Adaptive Weight Aggregation (FedAWA), a novel method that adaptively adjusts aggregation weights based on client vectors during the learning process. The client vector captures the direction of model updates, reflecting local data variations, and is used to optimize the aggregation weight without requiring additional datasets or violating privacy. By assigning higher aggregation weights to local models whose updates align closely with the global optimization direction, FedAWA enhances the stability and generalization of the global model. Extensive experiments under diverse scenarios demonstrate the superiority of our method, providing a promising solution to the challenges of data heterogeneity in federated learning.
****

FreeUV: Ground-Truth-Free Realistic Facial UV Texture Recovery via Cross-Assembly Inference Strategy
Xingchao Yang, Takafumi Taketomi, Yuki Endo, Yoshihiro Kanamori; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 326-337
Recovering high-quality 3D facial textures from single-view 2D images is a challenging task, especially under constraints of limited data and complex facial details such as makeup, wrinkles, and occlusions. In this paper, we introduce FreeUV, a novel ground-truth-free UV texture recovery framework that eliminates the need for annotated or synthetic UV data. FreeUV leverages pre-trained stable diffusion model alongside a Cross-Assembly inference strategy to fulfill this objective. In FreeUV, separate networks are trained independently to focus on realistic appearance and structural consistency, and these networks are combined during inference to generate coherent textures. Our approach accurately captures intricate facial features and demonstrates robust performance across diverse poses and occlusions. Extensive experiments validate FreeUV's effectiveness, with results surpassing state-of-the-art methods in both quantitative and qualitative metrics. Additionally, FreeUV enables new applications, including local editing, facial feature interpolation, and multi-view texture recovery. By reducing data requirements, FreeUV offers a scalable solution for generating high-fidelity 3D facial textures suitable for real-world scenarios.
****

HarmonySet: A Comprehensive Dataset for Understanding Video-Music Semantic Alignment and Temporal Synchronization
Zitang Zhou, Ke Mei, Yu Lu, Tianyi Wang, Fengyun Rao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3152-3162
This paper introduces HarmonySet, a comprehensive dataset designed to advance video-music understanding. HarmonySet consists of 48,328 diverse video-music pairs, annotated with detailed information on rhythmic synchronization, emotional alignment, thematic coherence, and cultural relevance. We propose a multi-step human-machine collaborative framework for efficient annotation, combining human insights with machine-generated descriptions to identify key transitions and assess alignment across multiple dimensions. Additionally, we introduce a novel evaluation framework with tasks and metrics to assess the multi-dimensional alignment of video and music, including rhythm, emotion, theme, and cultural context. Our extensive experiments demonstrate that HarmonySet, along with the proposed evaluation framework, significantly improves the ability of multimodal models to capture and analyze the intricate relationships between video and music. Project page

: https://harmonyset.github.io/.
****

Rethinking Diffusion for Text-Driven Human Motion Generation: Redundant Representations, Evaluation, and Masked Autoregression
Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, Huaizu Jiang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27859-27871
Since 2023, Vector Quantization (VQ)-based discrete generation methods have rapidly dominated human motion generation, primarily surpassing diffusion-based continuous generation methods in standard performance metrics. However, VQ-based methods have inherent limitations. Representing continuous motion data as limited discrete tokens leads to inevitable information loss, reduces the diversity of generated motions, and restricts their ability to function effectively as motion priors or generation guidance. In contrast, the continuous space generation nature of diffusion-based methods makes them well-suited to address these limitations and with even potential for model scalability. In this work, we systematically investigate why current VQ-based methods perform well and explore the limitations of existing diffusion-based methods from the perspective of motion data representation and distribution. Drawing on these insights, we preserve the inherent strengths of a diffusion-based human motion generation model and gradually optimize it with inspiration from VQ-based approaches. Our approach introduces a human motion diffusion model enabled to perform masked autoregression, optimized with a reformed data representation and distribution. Additionally, we propose a more robust evaluation method to assess different approaches. Extensive experiments on various datasets demonstrate our method outperforms previous methods and achieves state-of-the-art performances.
****

StyleMaster: Stylize Your Video with Artistic Generation and Translation
Zixuan Ye, Huijuan Huang, Xintao Wang, Pengfei Wan, Di Zhang, Wenhan Luo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2630-2640
Style control has been popular in video generation models. Existing methods often generate videos far from the given style, cause content leakage, and struggle to transfer one video to the desired style. Our first observation is that the style extraction stage matters, whereas existing methods emphasize global style but ignore local textures. In order to bring texture features while preventing content leakage, we filter content-related patches while retaining style ones based on prompt-patch similarity; for global style extraction, we generate a paired style dataset through model illusion to facilitate contrastive learning, which greatly enhances the absolute style consistency. Moreover, to fill in the image-to-video gap, we train a lightweight motion adapter on still videos, which implicitly enhances stylization extent, and enables our image-trained model to be seamlessly applied to videos. Benefited from these efforts, our approach, StyleMaster, not only achieves significant improvement in both style resemblance and temporal coherence, but also can easily generalize to video style transfer with a gray tile ControlNet. Extensive experiments and visualizations demonstrate that StyleMaster significantly outperforms competitors, effectively generating high-quality stylized videos that align with textual content and closely resemble the style of reference images.
****

Unsupervised Continual Domain Shift Learning with Multi-Prototype Modeling
Haopeng Sun, Yingwei Zhang, Lumin Xu, Sheng Jin, Ping Luo, Chen Qian, Wentao Liu, Yiqiang Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10131-10141
In real-world applications, deep neural networks may encounter constantly changing environments, where the test data originates from continually shifting unlabeled target domains. This problem, known as Unsupervised Continual Domain Shift Learning (UCDSL), poses practical difficulties. Existing methods for UCDSL aim to learn domain-invariant representations for all target domains. However, due to the existence of adaptivity gap, the invariant representation may theoretically

lead to large joint errors. To overcome the limitation, we propose a novel UCDSL method, called Multi-Prototype Modeling (MPM). Our model comprises two key components: (1) Multi-Prototype Learning (MPL) for acquiring domain-specific representations using multiple domain-specific prototypes. MPL achieves domain-specific error minimization instead of enforcing feature alignment across different domains. (2) Bi-Level Graph Enhancer (BiGE) for enhancing domain-level and category-level representations, resulting in more accurate predictions. We provide theoretical and empirical analysis to demonstrate the effectiveness of our proposed method. We evaluate our approach on multiple benchmark datasets and show that our model surpasses state-of-the-art methods across all datasets, highlighting its effectiveness and robustness in handling unsupervised continual domain shift learning. Codes will be publicly accessible.

****

OmniGuard: Hybrid Manipulation Localization via Augmented Versatile Deep Image Watermarking

Xuanyu Zhang, Zecheng Tang, Zhipei Xu, Runyi Li, Youmin Xu, Bin Chen, Feng Gao, Jian Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3008-3018

With the rapid growth of generative AI and its widespread application in image editing, new risks have emerged regarding the authenticity and integrity of digital content. Existing versatile watermarking approaches suffer from trade-offs between tamper localization precision and visual quality. Constrained by the limited flexibility of previous framework, their localized watermark must remain fixed across all images. Under AIGC-editing, their copyright extraction accuracy is also unsatisfactory. To address these challenges, we propose OmniGuard, a novel augmented versatile watermarking approach that integrates proactive embedding with passive, blind extraction for robust copyright protection and tamper localization. OmniGuard employs a hybrid forensic framework that enables flexible localization watermark selection and introduces a degradation-aware tamper extraction network for precise localization under challenging conditions. Additionally, a lightweight AIGC-editing simulation layer is designed to enhance robustness across global and local editing. Extensive experiments show that OmniGuard achieves superior fidelity, robustness, and flexibility. Compared to the recent state-of-the-art approach EditGuard, our method outperforms it by 4.25dB in PSNR of the container image, 20.7% in F1-Score under noisy conditions, and 14.8% in average bit accuracy.

****

Open-Canopy: Towards Very High Resolution Forest Monitoring

Fajwel Fogel, Yohann Perron, Nikola Besic, Laurent Saint-André, Agnès Pellissier-Tanon, Martin Schwartz, Thomas Boudras, Ibrahim Fayad, Alexandre d'Aspremont, Loic Landrieu, Philippe Ciais; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1395-1406

Estimating canopy height and its changes at meter resolution from satellite imagery is a significant challenge in computer vision with critical environmental applications. However, the lack of open-access datasets at this resolution hinders the reproducibility and evaluation of models. We introduce Open-Canopy, the first open-access, country-scale benchmark for very high-resolution (1.5 m) canopy height estimation, covering over 87,000 km2 across France with 1.5 m resolution satellite imagery and aerial LiDAR data. Additionally, we present Open-Canopy-, a benchmark for canopy height change detection between images from different years at tree level--a challenging task for current computer vision models. We evaluate state-of-the-art architectures on these benchmarks, highlighting significant challenges and opportunities for improvement. Our datasets and code are publicly available at https://github.com/fajwel/Open-Canopy.

****

ClearSight: Visual Signal Enhancement for Object Hallucination Mitigation in Multimodal Large Language Models

Hao Yin, Guangzong Si, Zilei Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14625-14634

Contrastive decoding strategies are widely used to mitigate object hallucination

s in multimodal large language models (MLLMs). By reducing over-reliance on lang
uage priors, these strategies ensure that generated content remains closely grou
nded in visual inputs, producing contextually accurate outputs. Since contrastiv
e decoding requires no additional training or external tools, it offers both com
putational efficiency and versatility, making it highly attractive. However, the
se methods present two main limitations: (1) bluntly suppressing language priors
 can compromise coherence and accuracy of generated content, and (2) processing
contrastive inputs adds computational load, significantly slowing inference spee
d. To address these challenges, we propose Visual Amplification Fusion (VAF), a
plug-and-play technique that enhances attention to visual signals within the mod
el's middle layers, where modality fusion predominantly occurs. This approach en
ables more effective capture of visual features, reducing the model's bias towar
d language modality. Experimental results demonstrate that VAF significantly red
uces hallucinations across various MLLMs without affecting inference speed, whil
e maintaining coherence and accuracy in generated outputs.
****

Stretching Each Dollar: Diffusion Training from Scratch on a Micro-Budget
Vikash Sehwag, Xianghao Kong, Jingtao Li, Michael Spranger, Lingjuan Lyu; Procee
dings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp
. 28596-28608
As scaling laws in generative AI push performance, they simultaneously concentra
te the development of these models among actors with large computational resourc
es. With a focus on text-to-image (T2I) generative models, we aim to unlock this
 bottleneck by demonstrating very low-cost training of large-scale T2I diffusion
 transformer models. As the computational cost of transformers increases with th
e number of patches in each image, we propose randomly masking up to 75% of the
image patches during training. We propose a deferred masking strategy that prepr
ocesses all patches using a patch-mixer before masking, thus significantly reduc
ing the performance degradation with masking, making it superior to model downsc
aling in reducing computational cost. We also incorporate the latest improvement
s in transformer architecture, such as the use of mixture-of-experts layers, to
improve performance and further identify the critical benefit of using synthetic
 images in micro-budget training. Finally, using only 37M publicly available rea
l and synthetic images, we train a 1.16 billion parameter sparse transformer wit
h only 1,890 USD economical cost and achieve a 12.7 FID in zero-shot generation
on the COCO dataset. Notably, our model achieves competitive performance across
both automated and human-centric evaluations, as well as high-quality generation
s, while incurring 118xlower costs than Stable Diffusion models and 14xlower cos
ts than the current state-of-the-art approach, which costs \28,400. We also furt
her investigate the influence of synthetic images on performance and demonstrate
 that micro-budget training on only synthetic images is sufficient for achieving
 high-quality data generation. Our end-to-end training pipeline and model checkp
oints are available at https://github.com/SonyResearch/micro_diffusion.
****

Guiding Human-Object Interactions with Rich Geometry and Relations
Mengqing Xue, Yifei Liu, Ling Guo, Shaoli Huang, Changxing Ding; Proceedings of
the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22714-2
2723
Human-object interaction (HOI) synthesis is crucial for creating immersive and r
ealistic experiences for applications such as virtual reality. Existing methods
often rely on simplified object representations, such as the object's centroid o
r the nearest point to a human, to achieve physically plausible motions. However
, these approaches may overlook geometric complexity, resulting in suboptimal in
teraction fidelity. To address this limitation, we introduce ROG, a novel diffus
ion-based framework that models the spatiotemporal relationships inherent in HOI
s with rich geometric detail. For efficient object representation, we select bou
ndary-focused and fine-detail key points from the object mesh, ensuring a compre
hensive depiction of the object's geometry. This representation is used to const
ruct an interactive distance field (IDF), capturing the robust HOI dynamics. Fur
thermore, we develop a diffusion-based relation model that integrates spatial an

d temporal attention mechanisms, enabling a better understanding of intricate HOI relationships. This relation model refines the generated motion's IDF, guiding the motion generation process to produce relation-aware and semantically aligned movements. Experimental evaluations demonstrate that ROG significantly outperforms state-of-the-art methods in the realism and semantic accuracy of synthesized HOIs. This paper's code will be released.
****

TacoDepth: Towards Efficient Radar-Camera Depth Estimation with One-stage Fusion
Yiran Wang, Jiaqi Li, Chaoyi Hong, Ruibo Li, Liusheng Sun, Xiao Song, Zhe Wang, Zhiguo Cao, Guosheng Lin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10523-10533
Radar-Camera depth estimation aims to predict dense and accurate metric depth by fusing input images and Radar data. Model efficiency is crucial for this task in pursuit of real-time processing on autonomous vehicles and robotic platforms. However, due to the sparsity of Radar returns, the prevailing methods adopt multi-stage frameworks with intermediate quasi-dense depth, which are time-consuming and not robust. To address these challenges, we propose TacoDepth, an efficient and accurate Radar-Camera depth estimation model with one-stage fusion. Specifically, the graph-based Radar structure extractor and the pyramid-based Radar fusion module are designed to capture and integrate the graph structures of Radar point clouds, delivering superior model efficiency and robustness without relying on the intermediate depth results. Moreover, TacoDepth can be flexible for different inference modes, providing a better balance of speed and accuracy. Extensive experiments are conducted to demonstrate the efficacy of our method. Compared with the previous state-of-the-art approach, TacoDepth improves depth accuracy and processing speed by 12.8% and 91.8%. Our work provides a new perspective on efficient Radar-Camera depth estimation.
****

Physical Plausibility-aware Trajectory Prediction via Locomotion Embodiment
Hiromu Taketsugu, Takeru Oba, Takahiro Maeda, Shohei Nobuhara, Norimichi Ukita; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12324-12334
Humans can predict future human trajectories even from momentary observations by using human pose-related cues. However, previous Human Trajectory Prediction (HTP) methods leverage the pose cues implicitly, resulting in implausible predictions. To address this, we propose Locomotion Embodiment, a framework that explicitly evaluates the physical plausibility of the predicted trajectory by locomotion generation under the laws of physics. While the plausibility of locomotion is learned with an indifferentiable physics simulator, it is replaced by our differentiable Locomotion Value function to train an HTP network in a data-driven manner. In particular, our proposed Embodied Locomotion loss is beneficial for efficiently training a stochastic HTP network using multiple heads. Furthermore, the Locomotion Value filter is proposed to filter out implausible trajectories at inference. Experiments demonstrate that our method enhances even the state-of-the-art HTP methods across diverse datasets and problem settings. Our code is available at: https://github.com/ImIntheMiddle/EmLoco.
****

CADDreamer: CAD Object Generation from Single-view Images
Yuan Li, Cheng Lin, Yuan Liu, Xiaoxiao Long, Chenxu Zhang, Ningna Wang, Xin Li, Wenping Wang, Xiaohu Guo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21448-21457
The field of diffusion-based 3D generation has experienced tremendous progress in recent times. However, existing 3D generative models often produce overly dense and unstructured meshes, which are in stark contrast to the compact, structured and clear-edged CAD models created by human modelers. We introduce CADDreamer, a novel method for generating CAD objects from a single image. This method proposes a primitive-aware multi-view diffusion model, which perceives both local geometry and high-level structural semantics during the generation process. We encode primitive semantics into the color domain, and enforce the strong priors in pre-trained diffusion models to align with the well-defined primitives. As a res

ult, we can infer multi-view normal maps and semantic maps from a single image, thereby reconstructing a mesh with primitive labels. Correspondingly, we propose a set of fitting and optimization methods to deal with the inevitable noise and distortion in generated primitives, ultimately producing a complete and seamless Boundary Representation (B-rep) of a Computer-Aided Design (CAD) model. Experimental results demonstrate that our method can effectively recover high-quality CAD objects from single-view images. Compared to existing 3D generation methods, the models produced by CADDreamer are compact in representation, clear in structure, sharp in boundaries, and watertight in topology.
****

Vision-Language Model IP Protection via Prompt-based Learning
Lianyu Wang, Meng Wang, Huazhu Fu, Daoqiang Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9497-9506
Vision-language models (VLMs) like CLIP (Contrastive Language-Image Pre-Training) have seen remarkable success in visual recognition, highlighting the increasing need to safeguard the intellectual property (IP) of well-trained models. Effective IP protection extends beyond ensuring authorized usage; it also necessitates restricting model deployment to authorized data domains, particularly when the model is fine-tuned for specific target domains. However, current IP protection methods often rely solely on the visual backbone, which may lack sufficient semantic richness. To bridge this gap, we introduce IP-CLIP, a lightweight IP protection strategy tailored to CLIP, employing a prompt-based learning approach. By leveraging the frozen visual backbone of CLIP, we extract both image style and content information, incorporating them into the learning of IP prompt. This strategy acts as a robust barrier, effectively preventing the unauthorized transfer of features from authorized domains to unauthorized ones. Additionally, we propose a style-enhancement branch that constructs feature banks for both authorized and unauthorized domains. This branch integrates self-enhanced and cross-domain features, further strengthening IP-CLIP's capability to block features from unauthorized domains. Finally, we present new three metrics designed to better balance the performance degradation of authorized and unauthorized domains. Comprehensive experiments in various scenarios demonstrate its promising potential for application in IP protection tasks for VLMs.
****

Where's the Liability in the Generative Era? Recovery-based Black-Box Detection of AI-Generated Content
Haoyue Bai, Yiyou Sun, Wei Cheng, Haifeng Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28821-28830
The recent proliferation of photorealistic images created by generative models has sparked both excitement and concern, as these images are increasingly indistinguishable from real ones to the human eye. While offering new creative and commercial possibilities, the potential for misuse, such as in misinformation and fraud, highlights the need for effective detection methods. Current detection approaches often rely on access to model weights or require extensive collections of real image datasets, limiting their scalability and practical application in real-world scenarios. In this work, we introduce a novel black-box detection framework that requires only API access, sidestepping the need for model weights or large auxiliary datasets. Our approach leverages a corrupt-and-recover strategy: by masking part of an image and assessing the model's ability to reconstruct it, we measure the likelihood that the image was generated by the model itself. For black-box models that do not support masked-image inputs, we incorporate a cost-efficient surrogate model trained to align with the target model's distribution, enhancing detection capability. Our framework demonstrates strong performance, outperforming baseline methods by 4.31% in mean average precision across eight diffusion model variant datasets.
****

Kiss3DGen: Repurposing Image Diffusion Models for 3D Asset Generation
Jiantao Lin, Xin Yang, Meixi Chen, Yingjie Xu, Dongyu Yan, Leyi Wu, Xinli Xu, Lie Xu, Shunsi Zhang, Ying-Cong Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5870-5880

Diffusion models have achieved great success in generating 2D images. However, t
he quality and generalizability of 3D content generation remain limited. State-o
f-the-art methods often require large-scale 3D assets for training, which are ch
allenging to collect. In this work, we introduce Kiss3DGen (Keep It Simple and S
traightforward in 3D Generation), an efficient framework for generating, editing
, and enhancing 3D objects by repurposing a well-trained 2D image diffusion mode
l for 3D generation. Specifically, we fine-tune a diffusion model to generate "3
D Bundle Image", a tiled representation composed of multi-view images and their
corresponding normal maps. The normal maps are then used to reconstruct a 3D mes
h, and the multi-view images provide texture mapping, resulting in a complete 3D
 model. This simple method effectively transforms the 3D generation problem into
 a 2D image generation task, maximizing the utilization of knowledge in pretrain
ed diffusion models. Furthermore, we demonstrate that our Kiss3DGen model is com
patible with various diffusion model techniques, enabling advanced features such
 as 3D editing, mesh and texture enhancement, etc. Through extensive experiments
, we demonstrate the effectiveness of our approach, showcasing its ability to pr
oduce high-quality 3D models efficiently.
****

DiTASK: Multi-Task Fine-Tuning with Diffeomorphic Transformations
Krishna Sri Ipsit Mantri, Carola-Bibiane Schönlieb, Bruno Ribeiro, Chaim Baskin,
 Moshe Eliasof; Proceedings of the Computer Vision and Pattern Recognition Confe
rence (CVPR), 2025, pp. 25218-25229
Pre-trained Vision Transformers now serve as powerful tools for computer vision.
 Yet, efficiently adapting them for multiple tasks remains a challenge that aris
es from the need to modify the rich hidden representations encoded by the learne
d weight matrices, without inducing interference between tasks. Current paramete
r-efficient methods like LoRA, which apply low-rank updates, force tasks to comp
ete within constrained subspaces, ultimately degrading performance. We introduce
 DiTASK, a novel Diffeomorphic Multi-Task Fine-Tuning approach that maintains pr
e-trained representations by preserving weight matrix singular vectors, while en
abling task-specific adaptations through neural diffeomorphic transformations of
 the singular values. By following this approach, DiTASK enables both shared and
 task-specific feature modulations with minimal added parameters. Our theoretic
al analysis shows that DiTASK achieves full-rank updates during optimization, pr
eserving the geometric structure of pre-trained features, and establishing a new
 paradigm for efficient multi-task learning (MTL). Our experiments on PASCAL MTL
 and NYUD show that DiTASK achieves state-of-the-art performance across four den
se prediction tasks, using 75% fewer parameters than existing methods.
****

OW-OVD: Unified Open World and Open Vocabulary Object Detection
Xing Xi, Yangyang Huang, Ronghua Luo, Yu Qiu; Proceedings of the Computer Vision
 and Pattern Recognition Conference (CVPR), 2025, pp. 25454-25464
Open world perception expands traditional closed-set frameworks, which assume a
predefined set of known categories, to encompass dynamic real-world environments
. Open World Object Detection (OWOD) and Open Vocabulary Object Detection (OVD)
are two main research directions, each addressing unique challenges in dynamic e
nvironments. However, existing studies often focus on only one of these tasks, l
eaving the combined challenges of OWOD and OVD largely underexplored. In this pa
per, we propose a novel detector, OW-OVD, which inherits the zero-shot generaliz
ation capability of OVD detectors while incorporating the ability to actively de
tect unknown objects and progressively optimize performance through incremental
learning, as seen in OWOD detectors. To achieve this, we start with a standard O
VD detector and adapt it for OWOD tasks. For attribute selection, we propose the
 Visual Similarity Attribute Selection (VSAS) method, which identifies the most
generalizable attributes by computing similarity distributions across annotated
and unannotated regions. Additionally, to ensure the diversity of attributes, we
 incorporate a similarity constraint in the iterative process. Finally, to prese
rve the standard inference process of OVD, we propose the Hybrid Attribute-Uncer
tainty Fusion (HAUF) method. This method combines attribute similarity with know
n class uncertainty to infer the likelihood of an object belonging to an unknown

class. We validated the effectiveness of OW-OVD through evaluations on two OWOD benchmarks, M-OWODB and S-OWODB. The results demonstrate that OW-OVD outperforms existing state-of-the-art models, achieving a +15.3 improvement in unknown object recall (U-Recall) and a +15.5 increase in unknown class average precision (U-mAP). Our code is available at: https://github.com/xxyzll/OW_OVD.
****

Improving Diffusion Inverse Problem Solving with Decoupled Noise Annealing
Bingliang Zhang, Wenda Chu, Julius Berner, Chenlin Meng, Anima Anandkumar, Yang Song; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20895-20905
Diffusion models have recently achieved success in solving Bayesian inverse problems with learned data priors. Current methods build on top of the diffusion sampling process, where each denoising step makes small modifications to samples from the previous step. However, this process struggles to correct errors from earlier sampling steps, leading to worse performance in complicated nonlinear inverse problems, such as phase retrieval. To address this challenge, we propose a new method called Decoupled Annealing Posterior Sampling (DAPS) that relies on a novel noise annealing process. Specifically, we decouple consecutive steps in a diffusion sampling trajectory, allowing them to vary considerably from one another while ensuring their time-marginals anneal to the true posterior as we reduce noise levels. This approach enables the exploration of a larger solution space, improving the success rate for accurate reconstructions. We demonstrate that DAPS significantly improves sample quality and stability across multiple image restoration tasks, particularly in complicated nonlinear inverse problems.
****

AvatarArtist: Open-Domain 4D Avatarization
Hongyu Liu, Xuan Wang, Ziyu Wan, Yue Ma, Jingye Chen, Yanbo Fan, Yujun Shen, Yibing Song, Qifeng Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10758-10769
This work focuses on open-domain 4D avatarization, with the purpose of creating a 4D avatar from a portrait image in an arbitrary style. We select parametric triplanes as the intermediate 4D representation, and propose a practical training paradigm that takes advantage of both generative adversarial networks (GANs) and diffusion models. Our design stems from the observation that 4D GANs excel at bridging images and triplanes without supervision yet usually face challenges in handling diverse data distributions. A robust 2D diffusion prior emerges as the solution, assisting the GAN in transferring its expertise across various domains. The synergy between these experts permits the construction of a multi-domain image-triplane dataset, which drives the development of a general 4D avatar creator. Extensive experiments suggest that our model, termed AvatarArtist, is capable of producing high-quality 4D avatars with strong robustness to various source image domains. The code, the data, and the models will be made publicly available to facilitate future studies.
****

DesignDiffusion: High-Quality Text-to-Design Image Generation with Diffusion Models
Zhendong Wang, Jianmin Bao, Shuyang Gu, Dong Chen, Wengang Zhou, Houqiang Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20906-20915
In this paper, we present DesignDiffusion, a simple yet effective framework for the novel task of synthesizing design images from textual descriptions. A primary challenge lies in generating accurate and style-consistent textual and visual content. Existing works in a related task of visual text generation often focus on generating text within given specific regions, which limits the creativity of generation models, resulting in style or color inconsistencies between textual and visual elements if applied to design image generation. To address this issue, we propose an end-to-end, one-stage diffusion-based framework that avoids intricate components like position and layout modeling. Specifically, the proposed framework directly synthesizes textual and visual design elements from user prompts. It utilizes a distinctive character embedding derived from the visual text t

o enhance the input prompt, along with a character localization loss for enhanced supervision during text generation. Furthermore, we employ a self-play Direct Preference Optimization fine-tuning strategy to improve the quality and accuracy of the synthesized visual text. Extensive experiments demonstrate that DesignDiffusion achieves state-of-the-art performance in design image generation.
****

Using Powerful Prior Knowledge of Diffusion Model in Deep Unfolding Networks for Image Compressive Sensing
Chen Liao, Yan Shen, Dan Li, Zhongli Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18000-18010
Recently, Deep Unfolding Networks (DUNs) have achieved impressive reconstruction quality in the field of image Compressive Sensing (CS) by unfolding iterative optimization algorithms into neural networks. The reconstruction quality of DUNs depends on the learned prior knowledge, so introducing stronger prior knowledge can further improve reconstruction quality. On the other hand, pre-trained diffusion models contain powerful prior knowledge and have a solid theoretical foundation and strong scalability, but it requires a large number of iterative steps to achieve reconstruction. In this paper, we propose to use the powerful prior knowledge of pre-trained diffusion model in DUNs to achieve high-quality reconstruction with less steps for image CS. Specifically, we first design an iterative optimization algorithm named Diffusion Message Passing (DMP), which embeds a pre-trained diffusion model into each iteration process of DMP. Then, we deeply unfold the DMP algorithm into a neural network named DMP-DUN. The proposed DMP-DUN can use lightweight neural networks to achieve mapping from measurement data to the intermediate steps of the reverse diffusion process and directly approximate the divergence of the diffusion model, thereby further improving reconstruction efficiency. Extensive experiments show that our proposed DMP-DUN achieves state-of-the-art performance and requires at least only 2 steps to reconstruct the image. Codes are available at https://github.com/FengodChen/DMP-DUN-CVPR2025.
****

Koala-36M: A Large-scale Video Dataset Improving Consistency between Fine-grained Conditions and Video Content
Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, Fei Yang, Pengfei Wan, Di Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8428-8437
With the continuous progress of visual generation technologies, the scale of video datasets has grown exponentially. The quality of these datasets plays a pivotal role in the performance of video generation models. We assert that temporal splitting, detailed captions, and video quality filtering are three crucial determinants of dataset quality. However, existing datasets exhibit various limitations in these areas. To address these challenges, we introduce Koala-36M, a large-scale, high-quality video dataset featuring accurate temporal splitting, detailed captions, and superior video quality. The essence of our approach lies in improving the consistency between fine-grained conditions and video content. Specifically, we employ a linear classifier on probability distributions to enhance the accuracy of transition detection, ensuring better temporal consistency. We then provide structured captions for the splitted videos, with an average length of 200 words, to improve text-video alignment. Additionally, we develop a Video Training Suitability Score (VTSS) that integrates multiple sub-metrics, allowing us to filter high-quality videos from the original corpus. Finally, we incorporate several metrics into the training process of the generation model, further refining the fine-grained conditions. Our experiments demonstrate the effectiveness of our data processing pipeline and the quality of the proposed Koala-36M dataset. Our dataset and code have been released at https://koala36m.github.io/.
****

VASparse: Towards Efficient Visual Hallucination Mitigation via Visual-Aware Token Sparsification
Xianwei Zhuang, Zhihong Zhu, Yuxin Xie, Liming Liang, Yuexian Zou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4189-

4199
Large Vision-Language Models (LVLMs) may produce outputs that are unfaithful to reality, also known as visual hallucinations (VH), which significantly impedes their real-world usage. To alleviate VH, various decoding strategies have been proposed to enhance visual information. However, many of these methods may require secondary decoding and rollback, which significantly reduces inference speed. In this work, we propose an efficient plug-and-play decoding algorithm via Visual-Aware Sparsification (VASparse) from the perspective of token sparsity for mitigating VH. VASparse is inspired by empirical observations: (1) the sparse activation of attention in LVLMs, and (2) visual-agnostic tokens sparsification exacerbates VH. Based on these insights, we propose a novel token sparsification strategy that balances efficiency and trustworthiness. Specifically, VASparse implements a visual-aware token selection strategy during decoding to reduce redundant tokens while preserving visual context effectively. Additionally, we innovatively introduce a sparse-based visual contrastive decoding method to recalibrate the distribution of hallucinated outputs without the time overhead associated with secondary decoding. Subsequently, VASparse recalibrates attention scores to penalize attention sinking of LVLMs towards text tokens. Extensive experiments across four popular benchmarks confirm the effectiveness of VASparse in mitigating VH across different LVLM families without requiring additional training or post-processing. Impressively, VASparse achieves state-of-the-art performance for mitigating VH while maintaining competitive decoding speed. Code is available at https://github.com/mengchuang123/VASparse-github.
****

SPARC: Score Prompting and Adaptive Fusion for Zero-Shot Multi-Label Recognition in Vision-Language Models
Kevin Miller, Aditya Gangrade, Samarth Mishra, Kate Saenko, Venkatesh Saligrama; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4313-4321
Zero-shot multi-label recognition (MLR) with Vision-Language Models (VLMs) faces significant challenges without training data, model tuning, or architectural modifications. Existing approaches require prompt tuning or architectural adaptations, limiting zero-shot applicability. Our work proposes a novel solution treating VLMs as black boxes, leveraging scores without training data or ground truth. We make two contributions. First, we find that VLM scores suffer from image- and prompt-specific biases, and that simple standardization is surprisingly effective at removing these and boosting MLR performance. And second, we introduce compound prompts grounded in realistic object combinations. Our analysis reveals "AND"/"OR" signal ambiguities that cause maximum compound scores to be surprisingly suboptimal compared to second-highest scores. We introduce an adaptive fusion method to address this issue. Our method enhances other zero-shot approaches, consistently improving their results. Experiments show superior mean Average Precision (mAP) compared to methods requiring training data, achieved through refined object ranking for robust zero-shot MLR. Code can be found at https://github.com/kjmillerCURIS/SPARC.
****

UniGoal: Towards Universal Zero-shot Goal-oriented Navigation
Hang Yin, Xiuwei Xu, Linqing Zhao, Ziwei Wang, Jie Zhou, Jiwen Lu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19057-19066
In this paper, we propose a general framework for universal zero-shot goal-oriented navigation. Existing zero-shot methods build inference framework upon large language models (LLM) for specific tasks, which differs a lot in overall pipeline and fails to generalize across different types of goal. Towards the aim of universal zero-shot navigation, we propose a uniform graph representation to unify different goals, including object category, instance image and text description. We also convert the observation of agent into an online maintained scene graph. With this consistent scene and goal representation, we preserve most structural information compared with pure text and are able to leverage LLM for explicit graph-based reasoning.Specifically, we conduct graph matching between the scene g

raph and goal graph at each time instant and propose different strategies to gen
erate long-term goal of exploration according to different matching states. The
agent first iteratively searches subgraph of goal when zero-matched. With partia
l matching, the agent then utilizes coordinate projection and anchor pair alignm
ent to infer the goal location. Finally scene graph correction and goal verifica
tion are applied for perfect matching. We also present a blacklist mechanism to
enable robust switch between stages.Extensive experiments on several benchmarks
show that our UniGoal achieves state-of-the-art zero-shot performance on three s
tudied navigation tasks with a single model, even outperforming task-specific ze
ro-shot methods and supervised universal methods.
****

Noise-Consistent Siamese-Diffusion for Medical Image Synthesis and Segmentation
Kunpeng Qiu, Zhiqiang Gao, Zhiying Zhou, Mingjie Sun, Yongxin Guo; Proceedings o
f the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15672
-15681
Deep learning has revolutionized medical image segmentation, yet its full potent
ial remains constrained by the paucity of annotated datasets. While diffusion mo
dels have emerged as a promising approach for generating synthetic image-mask pa
irs to augment these datasets, they paradoxically suffer from the same data scar
city challenges they aim to mitigate. Traditional mask-only models frequently yi
eld low-fidelity images due to their inability to adequately capture morphologic
al intricacies, which can critically compromise the robustness and reliability o
f segmentation models. To alleviate this limitation, we introduce Siamese-Diffus
ion, a novel dual-component model comprising Mask-Diffusion and Image-Diffusion.
 During training, a Noise Consistency Loss is introduced between these component
s to enhance the morphological fidelity of Mask-Diffusion in the parameter space
. During sampling, only Mask-Diffusion is used, ensuring diversity and scalabili
ty. Comprehensive experiments demonstrate the superiority of our method. Siamese
-Diffusion boosts SANet's mDice and mIoU by 3.6% and 4.4% on the Polyps, while U
Net improves by 1.52% and 1.64% on the ISIC2018.
****

DefectFill: Realistic Defect Generation with Inpainting Diffusion Model for Visu
al Inspection
Jaewoo Song, Daemin Park, Kanghyun Baek, Sangyub Lee, Jooyoung Choi, Eunji Kim,
Sungroh Yoon; Proceedings of the Computer Vision and Pattern Recognition Confere
nce (CVPR), 2025, pp. 18718-18727
Developing effective visual inspection models remains challenging due to the sca
rcity of defect data. While image generation models have been used to synthesize
 defect images, producing highly realistic defects remains difficult. We propose
 DefectFill, a novel method for realistic defect generation that requires only a
 few reference defect images. It leverages a fine-tuned inpainting diffusion mod
el, optimized with our custom loss functions incorporating defect, object, and a
ttention terms. It enables precise capture of detailed, localized defect feature
s and their seamless integration into defect-free objects. Additionally, our Low
-Fidelity Selection method further enhances the defect sample quality. Experimen
ts show that DefectFill generates high-quality defect images, enabling visual in
spection models to achieve state-of-the-art performance on the MVTec AD dataset.
****

Less is More: Efficient Image Vectorization with Adaptive Parameterization
Kaibo Zhao, Liang Bao, Yufei Li, Xu Su, Ke Zhang, Xiaotian Qiao; Proceedings of
the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18166-1
8175
Image vectorization aims to convert raster images to vector ones, allowing for e
asy scaling and editing.Existing works mainly rely on preset parameters (i.e., a
 fixed number of paths and control points), ignoring the complexity of the image
 and posing significant challenges to practical applications.We demonstrate that
 such an assumption is often incorrect, as the preset paths or control points ma
y be neither essential nor enough to achieve accurate and editable vectorization
 results.Based on this key insight, in this paper, we propose AdaVec, an efficie
nt image vectorization method with adaptive parametrization, where the paths and

control points can be adjusted dynamically based on the complexity of the input raster image.In particular, we first decompose the input raster image into a set of pure-colored layers that are aligned with human perception.For each layer with varying shape complexity, we propose a novel allocation mechanism to adaptively adjust the control point distribution.We further adopt a differentiable rendering process to compose and optimize the shape and color parameters of each layer iteratively.Extensive experiments demonstrate that AdaVec outperforms the baselines qualitatively and quantitatively, in terms of computational efficiency, vectorization accuracy, and editing flexibility.

****

FedMIA: An Effective Membership Inference Attack Exploiting "All for One" Principle in Federated Learning

Gongxi Zhu, Donghao Li, Hanlin Gu, Yuan Yao, Lixin Fan, Yuxing Han; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20643-20653

Federated Learning (FL) is a promising approach for training machine learning models on decentralized data while preserving privacy. However, privacy risks, particularly Membership Inference Attacks (MIAs), which aim to determine whether a specific data point belongs to a target client's training set, remain a significant concern. Existing methods for implementing MIAs in FL primarily analyze updates from the target client, focusing on metrics such as loss, gradient norm, and gradient difference. However, these methods fail to leverage updates from non-target clients, potentially underutilizing available information.In this paper, we first formulate a one-tailed likelihood-ratio hypothesis test based on the likelihood of updates from non-target clients. Building upon this formulation, we introduce a three-stage Membership Inference Attack (MIA) method, called FedMIA, which follows the "all for one"--leveraging updates from all clients across multiple communication rounds to enhance MIA effectiveness. Both theoretical analysis and extensive experimental results demonstrate that FedMIA outperforms existing MIAs in both classification and generative tasks. Additionally, it can be integrated as an extension to existing methods and is robust against various defense strategies, Non-IID data, and different federated structures. Our code is available in https://github.com/Liar-Mask/FedMIA.

****

Erase Diffusion: Empowering Object Removal Through Calibrating Diffusion Pathways

Yi Liu, Hao Zhou, Benlei Cui, Wenxiang Shang, Ran Lin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2418-2427

Erase inpainting, or object removal, aims to precisely remove target objects within masked regions while preserving the overall consistency of the surrounding content. Despite diffusion-based methods have made significant strides in the field of image inpainting, challenges remain regarding the emergence of unexpected objects or artifacts. We assert that the inexact diffusion pathways established by existing standard optimization paradigms constrain the efficacy of object removal. To tackle these challenges, we propose a novel Erase Diffusion, termed EraDiff, aimed at unleashing the potential power of standard diffusion in the context of object removal. In contrast to standard diffusion, the EraDiff adapts both the optimization paradigm and the network to improve the coherence and elimination of the erasure results.We first introduce a Chain-Rectifying Optimization (CRO) paradigm, a sophisticated diffusion process specifically designed to align with the objectives of erasure. This paradigm establishes innovative diffusion transition pathways that simulate the gradual elimination of objects during optimization, allowing the model to accurately capture the intent of object removal. Furthermore, to mitigate deviations caused by artifacts during the sampling pathways, we develop a simple yet effective Self-Rectifying Attention (SRA) mechanism. The SRA calibrates the sampling pathways by altering self-attention activation, allowing the model to effectively bypass artifacts while further enhancing the coherence of the generated content. With this design, our proposed EraDiff achieves state-of-the-art performance on the OpenImages V5 dataset and demonstrates significant superiority in real-world scenarios.

****
Prompt-CAM: Making Vision Transformers Interpretable for Fine-Grained Analysis
Arpita Chowdhury, Dipanjyoti Paul, Zheda Mai, Jianyang Gu, Ziheng Zhang, Kazi Sajeed Mehrab, Elizabeth G. Campolongo, Daniel Rubenstein, Charles V. Stewart, Anuj Karpatne, Tanya Berger-Wolf, Yu Su, Wei-Lun Chao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4375-4385
We present a simple approach to make pre-trained Vision Transformers (ViTs) interpretable for fine-grained analysis, aiming to identify and localize the traits that distinguish visually similar categories, such as bird species. Pre-trained ViTs, such as DINO, have demonstrated remarkable capabilities in extracting localized, discriminative features. However, saliency maps like Grad-CAM often fail to identify these traits, producing blurred, coarse heatmaps that highlight entire objects instead. We propose a novel approach, Prompt Class Attention Map (Prompt-CAM), to address this limitation. Prompt-CAM learns class-specific prompts for a pre-trained ViT and uses the corresponding outputs for classification. To correctly classify an image, the true-class prompt must attend to unique image patches not present in other classes' images (i.e., traits). As a result, the true class's multi-head attention maps reveal traits and their locations. Implementation-wise, Prompt-CAM is almost a "free lunch," requiring only a modification to the prediction head of Visual Prompt Tuning (VPT). This makes Prompt-CAM easy to train and apply, in stark contrast to other interpretable methods that require designing specific models and training processes. Extensive empirical studies on a dozen datasets from various domains (e.g., birds, fishes, insects, fungi, flowers, food, and cars) validate the superior interpretation capability of Prompt-CAM. The source code and demo are available at https://github.com/Imageomics/Prompt_CAM.
****
Instruction-based Image Manipulation by Watching How Things Move
Mingdeng Cao, Xuaner Zhang, Yinqiang Zheng, Zhihao Xia; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2704-2713
This paper introduces a novel dataset construction pipeline that samples pairs of frames from videos and uses multimodal large language models (MLLMs) to generate editing instructions for training instruction-based image manipulation models. Video frames inherently preserve the identity of subjects and scenes, ensuring consistent content preservation during editing. Additionally, video data captures diverse, natural dynamics--such as non-rigid subject motion and complex camera movements--that are difficult to model otherwise, making it an ideal source for scalable dataset construction. Using this approach, we create a new dataset to train InstructMove, a model capable of instruction-based complex manipulations that are difficult to achieve with synthetically generated datasets. Our model demonstrates state-of-the-art performance in tasks such as adjusting subject poses, rearranging elements, and altering camera perspectives.
****
DPFlow: Adaptive Optical Flow Estimation with a Dual-Pyramid Framework
Henrique Morimitsu, Xiaobin Zhu, Roberto M. Cesar, Xiangyang Ji, Xu-Cheng Yin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17810-17820
Optical flow estimation is essential for video processing tasks, such as restoration and action recognition. The quality of videos is constantly increasing, with current standards reaching 8K resolution. However, optical flow methods are usually designed for low resolution and do not generalize to large inputs due to their rigid architectures. They adopt downscaling or input tiling to reduce the input size, causing a loss of details and global information. There is also a lack of optical flow benchmarks to judge the actual performance of existing methods on high-resolution samples. Previous works only conducted qualitative high-resolution evaluations on hand-picked samples. This paper fills this gap in optical flow estimation in two ways. We propose DPFlow, an adaptive optical flow architecture capable of generalizing up to 8K resolution inputs while trained with only low-resolution samples. We also introduce Kubric-NK, a new benchmark for evaluating optical flow methods with input resolutions ranging from 1K to 8K. Our high

-resolution evaluation pushes the boundaries of existing methods and reveals new insights about their generalization capabilities. Extensive experimental results show that DPFlow achieves state-of-the-art results on the MPI-Sintel, KITTI 2015, Spring, and other high-resolution benchmarks. The code and dataset are available at https://github.com/hmorimitsu/ptlflow/tree/main/ptlflow/models/dpflow.
****

DocSAM: Unified Document Image Segmentation via Query Decomposition and Heterogeneous Mixed Learning
Xiao-Hui Li, Fei Yin, Cheng-Lin Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15021-15032
Document image segmentation is crucial in document analysis and recognition but remains challenging due to the heterogeneity of document formats and diverse segmentation tasks. Existing methods often treat these tasks separately, leading to limited generalization and resource wastage.This paper introduces DocSAM, a transformer-based unified framework for various document image segmentation tasks, including document layout analysis, multi-granularity text segmentation, and table structure recognition by modelling these tasks as a combination of instance and semantic segmentation.Specifically, DocSAM uses a Sentence BERT to map category names from each dataset into semantic queries of the same dimension as instance queries. These queries interact through attention mechanisms and are cross-attended with image features to predict instance and semantic segmentation masks. To predict instance categories, instance queries are dot-producted with semantic queries, and scores are normalized using softmax.As a result, DocSAM can be jointly trained on heterogeneous datasets, enhancing robustness and generalization while reducing computing and storage resources. Comprehensive evaluations show that DocSAM outperforms existing methods in accuracy, efficiency, and adaptability, highlighting its potential for advancing document image understanding and segmentation in various applications.
****

Ferret: An Efficient Online Continual Learning Framework under Varying Memory Constraints
Yuhao Zhou, Yuxin Tian, Jindi Lv, Mingjia Shi, Yuanxi Li, Qing Ye, Shuhao Zhang, Jiancheng Lv; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4850-4861
In the realm of high-frequency data streams, achieving real-time learning within varying memory constraints is paramount. This paper presents Ferret, a comprehensive framework designed to enhance online accuracy of Online Continual Learning (OCL) algorithms while dynamically adapting to varying memory budgets.Ferret employs a fine-grained pipeline parallelism strategy combined with an iterative gradient compensation algorithm, ensuring seamless handling of high-frequency data with minimal latency, and effectively counteracting the challenge of stale gradients in parallel training. To adapt to varying memory budgets, its automated model partitioning and pipeline planning optimizes performance regardless of memory limitations. Extensive experiments across 20 benchmarks and 5 integrated OCL algorithms show Ferret's remarkable efficiency, achieving up to 3.7x lower memory overhead to reach the same online accuracy compared to competing methods.Furthermore, Ferret consistently outperforms these methods across diverse memory budgets, underscoring its superior adaptability. These findings position Ferret as a premier solution for efficient and adaptive OCL framework in real-time environments.
****

Spatiotemporal Skip Guidance for Enhanced Video Diffusion Sampling
Junha Hyung, Kinam Kim, Susung Hong, Min-Jung Kim, Jaegul Choo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11006-11015
Diffusion models have emerged as a powerful tool for generating high-quality images, videos, and 3D content. While sampling guidance techniques like CFG improve quality, they reduce diversity and motion. Autoguidance mitigates these issues but demands extra weak model training, limiting its practicality for large-scale models. In this work, we introduce Spatiotemporal Skip Guidance (STG), a simple

training-free sampling guidance method for enhancing transformer-based video diffusion models. STG employs an implicit weak model via self-perturbation, avoiding the need for external models or additional training. By selectively skipping spatiotemporal layers, STG produces an aligned, degraded version of the original model to boost sample quality without compromising diversity or dynamic degree. Our contributions include: (1) introducing STG as an efficient, high-performing guidance technique for video diffusion models, (2) eliminating the need for auxiliary models by simulating a weak model through layer skipping, and (3) ensuring quality-enhanced guidance without compromising sample diversity or dynamics unlike CFG. For additional results, visit https://junhahyung.github.io/STGuidance.

****

VidComposition: Can MLLMs Analyze Compositions in Compiled Videos?

Yunlong Tang, Junjia Guo, Hang Hua, Susan Liang, Mingqian Feng, Xinyang Li, Rui Mao, Chao Huang, Jing Bi, Zeliang Zhang, Pooyan Fazli, Chenliang Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8490-8500

The advancement of Multimodal Large Language Models (MLLMs) has enabled significant progress in multimodal understanding, expanding their capacity to analyze video content. However, existing evaluation benchmarks for MLLMs primarily focus on abstract video comprehension, lacking a detailed assessment of their ability to understand video compositions, the nuanced interpretation of how visual elements combine and interact within highly compiled video contexts. We introduce VidComposition, a new benchmark specifically designed to evaluate the video composition understanding capabilities of MLLMs using carefully curated compiled videos and cinematic-level annotations.VidComposition includes 982 videos with 1706 multiple-choice questions, covering various compositional aspects such as camera movement, angle, shot size, narrative structure, character actions and emotions, etc. Our comprehensive evaluation of 33 open-source and proprietary MLLMs reveals a significant performance gap between human and model capabilities. This highlights the limitations of current MLLMs in understanding complex, compiled video compositions and offers insights into areas for further improvement. Our benchmark is publicly available at https://yunlong10.github.io/VidComposition/

****

SAR3D: Autoregressive 3D Object Generation and Understanding via Multi-scale 3D VQVAE

Yongwei Chen, Yushi Lan, Shangchen Zhou, Tengfei Wang, Xingang Pan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28371-28382

Autoregressive models have demonstrated remarkable success across various fields, from large language models (LLMs) to large multimodal models (LMMs) and 2D content generation, moving closer to artificial general intelligence (AGI). Despite these advances, applying autoregressive approaches to 3D object generation and understanding remains largely unexplored. This paper introduces Scale AutoRegressive 3D (SAR3D), a novel framework that leverages a multi-scale 3D vector-quantized variational autoencoder (VQVAE) to tokenize 3D objects for efficient autoregressive generation and detailed understanding. By predicting the next scale in a multi-scale latent representation instead of the next single token, SAR3D reduces generation time significantly, achieving fast 3D object generation in just 0.82 seconds on an A6000 GPU. Additionally, given the tokens enriched with hierarchical 3D-aware information, we finetune a pretrained LLM on them, enabling multimodal comprehension of 3D content.Our experiments show that SAR3D surpasses current 3D generation methods in both speed and quality and allows LLMs to interpret and caption 3D models comprehensively.

****

Dual-Interrelated Diffusion Model for Few-Shot Anomaly Image Generation

Ying Jin, Jinlong Peng, Qingdong He, Teng Hu, Jiafu Wu, Hao Chen, Haoxuan Wang, Wenbing Zhu, Mingmin Chi, Jun Liu, Yabiao Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30420-30429

The performance of anomaly inspection in industrial manufacturing is constrained by the scarcity of anomaly data. To overcome this challenge, researchers have s

tarted employing anomaly generation approaches to augment the anomaly dataset. H
owever, existing anomaly generation methods suffer from limited diversity in the
 generated anomalies and struggle to achieve a seamless blending of this anomaly
 with the original image. Moreover, the generated mask is usually not aligned wi
th the generated anomaly. In this paper, we overcome these challenges from a new
 perspective, simultaneously generating a pair of the overall image and the corr
esponding anomaly part. We propose DualAnoDiff, a novel diffusion-based few-shot
 anomaly image generation model, which can generate diverse and realistic anomal
y images by using a dual-interrelated diffusion model, where one of them is empl
oyed to generate the whole image while the other one generates the anomaly part.
 Moreover, we extract background and shape information to mitigate the distortio
n and blurriness phenomenon in few-shot image generation. Extensive experiments
 demonstrate the superiority of our proposed model over state-of-the-art methods
 in terms of diversity, realism and the accuracy of mask. Overall, our approach
significantly improves the performance of downstream anomaly inspection tasks, i
ncluding anomaly detection, anomaly localization, and anomaly classification tas
ks. Code will be made available.
****

ODE: Open-Set Evaluation of Hallucinations in Multimodal Large Language Models
Yahan Tu, Rui Hu, Jitao Sang; Proceedings of the Computer Vision and Pattern Rec
ognition Conference (CVPR), 2025, pp. 19836-19845
Hallucination poses a persistent challenge for multimodal large language models
(MLLMs). However, existing benchmarks for evaluating hallucinations are generall
y static, which may overlook the potential risk of data contamination. To addres
s this issue, we propose ODE, an open-set, dynamic protocol designed to evaluate
 object hallucinations in MLLMs at both the existence and attribute levels. ODE
employs a graph-based structure to represent real-world object concepts, their a
ttributes, and the distributional associations between them. This structure faci
litates the extraction of concept combinations based on diverse distributional c
riteria, generating varied samples for structured queries that evaluate hallucin
ations in both generative and discriminative tasks. Through the generation of ne
w samples, dynamic concept combinations, and varied distribution frequencies, OD
E mitigates the risk of data contamination and broadens the scope of evaluation.
 This protocol is applicable to both general and specialized scenarios, includin
g those with limited data. Experimental results demonstrate the effectiveness of
 our protocol, revealing that MLLMs exhibit higher hallucination rates when eval
uated with ODE-generated samples, which indicates potential data contamination.
Furthermore, these generated samples aid in analyzing hallucination patterns and
 fine-tuning models, offering an effective approach to mitigating hallucinations
 in MLLMs.
****

Self-Supervised Learning for Color Spike Camera Reconstruction
Yanchen Dong, Ruiqin Xiong, Xiaopeng Fan, Zhaofei Yu, Yonghong Tian, Tiejun Huan
g; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR),
 2025, pp. 6231-6240
Spike camera is a kind of neuromorphic camera with ultra-high temporal resolutio
n, which can capture dynamic scenes by continuously firing spike signals. To cap
ture color information, a color filter array (CFA) is employed on the sensor of
the spike camera, resulting in Bayer-pattern spike streams. How to restore high-
quality color images from the binary spike signals remains challenging. In this
paper, we propose a motion-guided reconstruction method for spike cameras with C
FA, utilizing color layout and estimated motion information. Specifically, we de
velop a joint motion estimation pipeline for the Bayer-pattern spike stream, exp
loiting the motion consistency of channels. We propose to estimate the missing p
ixels of each color channel according to temporally neighboring pixels of the co
rresponding color along the motion trajectory. As the spike signals are read out
 at discrete time points, there is quantization noise that impacts the image qua
lity. Thus, we analyze the correlation of the noise in spatial and temporal doma
ins and propose a self-supervised network utilizing a masked spike encoder to ha
ndle the noise. Experiments on real-world captured Bayer-pattern spike streams s

how that our method can restore color images with better visual quality, compare d with state-of-the-art methods. The source codes are available at https://githu b.com/csycdong/SSL-CSC.
****

Interactive Medical Image Analysis with Concept-based Similarity Reasoning
Ta Duc Huy, Sen Kim Tran, Phan Nguyen, Nguyen Hoang Tran, Tran Bao Sam, Anton va n den Hengel, Zhibin Liao, Johan W. Verjans, Minh-Son To, Vu Minh Hieu Phan; Pro ceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30797-30806
The ability to interpret and intervene model decisions is important for the adop tion of computer-aided diagnosis methods in clinical workflows. Recent concept-b ased methods link the model predictions with interpretable concepts and modify t heir activation scores to interact with the model. However, these concepts are a t the image level, which hinders the model from pinpointing the exact patches th e concepts are activated. Alternatively, prototype-based methods learn represent ations from training image patches and compare these with test image patches, us ing the similarity scores for final class prediction. However, interpreting the underlying concepts of these patches can be challenging and often necessitates p ost-hoc guesswork. To address this issue, this paper introduces the novel Concep t-based Similarity Reasoning network (CSR), which offers (i) patch-level prototy pe with intrinsic concept interpretation, and (ii) spatial interactivity. First, the proposed CSR provides localized explanation by grounding prototypes of each concept on image regions. Second, our model introduces novel spatial-level inte raction, allowing doctors to engage directly with specific image areas, making i t an intuitive and transparent tool for medical imaging. CSR improves upon prior state-of-the-art interpretable methods by up to 4.% across three biomedical dat asets.
****

From Elements to Design: A Layered Approach for Automatic Graphic Design Composi tion
Jiawei Lin, Shizhao Sun, Danqing Huang, Ting Liu, Ji Li, Jiang Bian; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 812 8-8137
In this work, we investigate automatic design composition from multimodal graphi c elements. Although recent studies have developed various generative models for graphic design, they usually face the following limitations: they only focus on certain subtasks and are far from achieving the design composition task; they d o not consider the hierarchical information of graphic designs during the genera tion process. To tackle these issues, we introduce the layered design principle into Large Multimodal Models (LMMs) and propose a novel approach, called LaDeCo, to accomplish this challenging task. Specifically, LaDeCo first performs layer planning for a given element set, dividing the input elements into different sem antic layers according to their contents. Based on the planning results, it subs equently predicts element attributes that control the design composition in a la yer-wise manner, and includes the rendered image of previously generated layers into the context. With this insightful design, LaDeCo decomposes the difficult t ask into smaller manageable steps, making the generation process smoother and cl earer. The experimental results demonstrate the effectiveness of LaDeCo in desig n composition. Furthermore, we show that LaDeCo enables some interesting applica tions in graphic design, such as resolution adjustment, design decoration, desig n variation, etc. In addition, it even outperforms the specialized models in som e design subtasks without any task-specific training.
****

h-Edit: Effective and Flexible Diffusion-Based Editing via Doob's h-Transform
Toan Nguyen, Kien Do, Duc Kieu, Thin Nguyen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28490-28501
We introduce a theoretical framework for diffusion-based image editing by formul ating it as a reverse-time bridge modeling problem. This approach modifies the b ackward process of a pretrained diffusion model to construct a bridge that conve rges to an implicit distribution associated with the editing target at time 0. B

uilding on this framework, we propose h-Edit, a novel editing method that utiliz es Doob's h-transform and Langevin Monte Carlo to decompose the update of an int ermediate edited sample into two components: a "reconstruction" term and an "edi ting" term. This decomposition provides flexibility, allowing the reconstruction term to be computed via existing inversion techniques and enabling the combinat ion of multiple editing terms to handle complex editing tasks. To our knowledge, h-Edit is the first training-free method capable of performing simultaneous tex t-guided and reward-model-based editing. Extensive experiments, both quantitativ e and qualitative, show that h-Edit outperforms state-of-the-art baselines in te rms of editing effectiveness and faithfulness.
****

Masking meets Supervision: A Strong Learning Alliance
Byeongho Heo, Taekyung Kim, Sangdoo Yun, Dongyoon Han; Proceedings of the Comput er Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20447-20457
Pre-training with random masked inputs has emerged as a novel trend in self-supe rvised training. However, supervised learning still faces a challenge in adoptin g masking augmentations, primarily due to unstable training. In this paper, we p ropose a novel way to involve masking augmentations dubbed Masked Sub-branch (Ma skSub). MaskSub consists of the main-branch and sub-branch, the latter being a p art of the former. The main-branch undergoes conventional training recipes, whil e the sub-branch merits intensive masking augmentations, during training. MaskSu b tackles the challenge by mitigating adverse effects through a relaxed loss fun ction similar to a self-distillation loss. Our analysis shows that MaskSub impro ves performance, with the training loss converging faster than in standard train ing, which suggests our method stabilizes the training process. We further valid ate MaskSub across diverse training scenarios and models, including DeiT-III tra ining, MAE finetuning, CLIP finetuning, BERT training, and hierarchical architec tures (ResNet and Swin Transformer). Our results show that MaskSub consistently achieves impressive performance gains across all the cases. MaskSub provides a p ractical and effective solution for introducing additional regularization under various training recipes. Code available at https://github.com/naver-ai/augsub
****

DI-PCG: Diffusion-based Efficient Inverse Procedural Content Generation for High -quality 3D Asset Creation
Wang Zhao, Yan-Pei Cao, Jiale Xu, Yuejiang Dong, Ying Shan; Proceedings of the C omputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11061-11072
Procedural Content Generation (PCG) is powerful in creating high-quality 3D cont ents, yet controlling it to produce desired shapes is difficult and often requir es extensive parameter tuning. Inverse Procedural Content Generation aims to aut omatically find the best parameters under the input condition. However, existing sampling-based and neural network-based methods still suffer from numerous samp le iterations or limited controllability. In this work, we present DI-PCG, a nov el and efficient method for Inverse PCG from general image conditions. At its co re is a lightweight diffusion transformer model, where PCG parameters are direct ly treated as the denoising target and the observed images as conditions to cont rol parameter generation. DI-PCG is efficient and effective. With only 7.6M netw ork parameters and 30 GPU hours to train, it demonstrates superior performance i n recovering parameters accurately, and generalizing well to in-the-wild images. Quantitative and qualitative experiment results validate the effectiveness of D I-PCG in inverse PCG and image-to-3D generation tasks. DI-PCG offers a promising approach for efficient inverse PCG and represents a valuable exploration step t owards a 3D generation path that models how to construct a 3D asset using parame tric models.
****

SALOVA: Segment-Augmented Long Video Assistant for Targeted Retrieval and Routin g in Long-Form Video Analysis
Junho Kim, Hyunjun Kim, Hosu Lee, Yong Man Ro; Proceedings of the Computer Visio n and Pattern Recognition Conference (CVPR), 2025, pp. 3352-3362
Despite advances in Large Multi-modal Models, applying them to long and untrimme d video content remains challenging due to limitations in context length and sub

stantial memory overhead. These constraints often lead to significant informatio n loss and reduced relevance in the model responses. With the exponential growth of video data across web platforms, understanding long-form video is crucial fo r advancing generalized intelligence. In this paper, we introduce SALOVA: Segmen t-Augmented LOng Video Assistant, a novel video-LLM framework designed to enhanc e the comprehension of lengthy video content through targeted retrieval process. We address two main challenges to achieve it: (i) We present the SceneWalk data set, a high-quality collection of 87.8K long videos, each densely captioned at t he segment level to enable models to capture scene continuity and maintain rich descriptive context. (ii) We develop robust architectural designs integrating dy namic routing mechanism and spatio-temporal projector to efficiently retrieve an d process relevant video segments based on user queries. Our framework mitigates the limitations of current video-LMMs by allowing for precise identification an d retrieval of relevant video segments in response to queries, thereby improving the contextual relevance of the generated responses. Through extensive experime nts, SALOVA demonstrates enhanced capability in processing complex long-form vid eos, showing significant capability to maintain contextual integrity across exte nded sequences.

****

Notes-guided MLLM Reasoning: Enhancing MLLM with Knowledge and Visual Notes for Visual Question Answering

Wenlong Fang, Qiaofeng Wu, Jing Chen, Yun Xue; Proceedings of the Computer Visio n and Pattern Recognition Conference (CVPR), 2025, pp. 19597-19607

The knowledge-based visual question answering (KB-VQA) task involves using exter nal knowledge about the image to assist reasoning. Building on the impressive pe rformance of multimodal large language model (MLLM), recent methods have commenc ed leveraging MLLM as an implicit knowledge base for reasoning. However, the dir ect employment of MLLM with raw external knowledge might result in reasoning err ors due to misdirected knowledge information. Additionally, MLLM may lack fine-grained perception of visual features, which can result in hallucinations during reasoning. To address these challenges, we propose Notes-guided MLLM Reasoning (NoteMR), a novel framework that guides MLLM in better reasoning by utilizing kn owledge notes and visual notes. Specifically, we initially obtain explicit knowl edge from an external knowledge base. Then, this explicit knowledge, combined wi th images, is used to assist the MLLM in generating knowledge notes. These notes are designed to filter explicit knowledge and identify relevant internal implic it knowledge within the MLLM. We then identify highly correlated regions between the images and knowledge notes, retaining them as image notes to enhance the mo del's fine-grained perception, thereby mitigating MLLM induced hallucinations. F inally, both notes are fed into the MLLM, enabling a more comprehensive understa nding of the image-question pair and enhancing the model's reasoning capabilitie s. Our method achieves state-of-the-art performance on the OK-VQA and A-OKVQA da tasets, demonstrating its robustness and effectiveness across diverse VQA scenar ios.

****

Are Spatial-Temporal Graph Convolution Networks for Human Action Recognition Ove r-Parameterized?

Jianyang Xie, Yitian Zhao, Yanda Meng, He Zhao, Anh Nguyen, Yalin Zheng; Proceed ings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24309-24319

Spatial-temporal graph convolutional networks (ST-GCNs) showcase impressive perf ormance in skeleton-based human action recognition (HAR). However, despite the d evelopment of numerous models, their recognition performance does not differ sig nificantly after aligning the input settings. With this observation, we hypothes ize that ST-GCNs are over-parameterized for HAR, a conjecture subsequently confi rmed through experiments employing the lottery ticket hypothesis. Additionally, a novel sparse ST-GCNs generator is proposed, which trains a sparse architecture from a randomly initialized dense network while maintaining comparable performa nce levels to the dense components. Moreover, we generate multi-level sparsity S T-GCNs by integrating sparse structures at various sparsity levels and demonstra

te that the assembled model yields a significant enhancement in HAR performance. Thorough experiments on four datasets, including NTU-RGB+D 60(120), Kinetics-400, and FineGYM, demonstrate that the proposed sparse ST-GCNs can achieve comparable performance to their dense components. Even with 95% fewer parameters, the sparse ST-GCNs exhibit a degradation of <1% in top-1 accuracy. Meanwhile, the multi-level sparsity ST-GCNs, which require only 66% of the parameters of the dense ST-GCNs, demonstrate an improvement of >1% in top-1 accuracy. The code is available at https://github.com/davelailai/Sparse-ST-GCN.
****

DA-VPT: Semantic-Guided Visual Prompt Tuning for Vision Transformers
Li Ren, Chen Chen, Liqiang Wang, Kien Hua; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4353-4363
Visual Prompt Tuning (VPT) has become a promising solution for Parameter-Efficient Fine-Tuning (PEFT) approach for Vision Transformer (ViT) models by partially fine-tuning learnable tokens while keeping most model parameters frozen. Recent research has explored modifying the connection structures of the prompts. However, the fundamental correlation and distribution between the prompts and image tokens remain unexplored. In this paper, we leverage metric learning techniques to investigate how the distribution of prompts affects fine-tuning performance. Specifically, we propose a novel framework, Distribution Aware Visual Prompt Tuning (DA-VPT), to guide the distributions of the prompts by learning the distance metric from their class-related semantic data. Our method demonstrates that the prompts can serve as an effective bridge to share semantic information between image patches and the class token. We extensively evaluated our approach on popular benchmarks in both recognition and segmentation tasks. The results demonstrate that our approach enables more effective and efficient fine-tuning of ViT models by leveraging semantic information to guide the learning of the prompts, leading to improved performance on various downstream vision tasks.
****

Towards Lossless Implicit Neural Representation via Bit Plane Decomposition
Woo Kyoung Han, Byeonghun Lee, Hyunmin Cho, Sunghoon Im, Kyong Hwan Jin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2269-2278
We quantify the upper bound on the size of the implicit neural representation (INR) model from a digital perspective. The upper bound of the model size increases exponentially as the required bit-precision increases. To this end, we present a bit-plane decomposition method that makes INR predict bit-planes, producing the same effect as reducing the upper bound of the model size. We validate our hypothesis that reducing the upper bound leads to faster convergence with constant model size. Our method achieves lossless representation in 2D image and audio fitting, even for high bit-depth signals, such as 16-bit, which was previously unachievable. We pioneered the presence of bit bias, which INR prioritizes as the most significant bit (MSB). We expand the application of the INR task to bit depth expansion, lossless image compression, and extreme network quantization. Our source code is available at https://github.com/WooKyoungHan/LosslessINR.
****

Spectral State Space Model for Rotation-Invariant Visual Representation Learning
Sahar Dastani, Ali Bahri, Moslem Yazdanpanah, Mehrdad Noori, David Osowiechi, Gustavo Adolfo Vargas Hakim, Farzad Beizaee, Milad Cheraghalikhani, Arnab Kumar Mondal, Herve Lombaert, Christian Desrosiers; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23881-23890
State Space Models (SSMs) have recently emerged as an alternative to Vision Transformers (ViTs) due to their unique ability of modeling global relationships with linear complexity. SSMs are specifically designed to capture spatially proximate relationships of image patches. However, they fail to identify relationships between conceptually related yet not adjacent patches. This limitation arises from the non-causal nature of image data, which lacks inherent directional relationships. Additionally, current vision-based SSMs are highly sensitive to transformations such as rotation. Their predefined scanning directions depend on the original image orientation, which can cause the model to produce inconsistent patch

-processing sequences after rotation.To address these limitations, we introduce Spectral VMamba, a novel approach that effectively captures the global structure within an image by leveraging spectral information derived from the graph Laplacian of image patches. Through spectral decomposition, our approach encodes patch relationships independently of image orientation, achieving rotation invariance with the aid of our Rotational Feature Normalizer (RFN) module. Our experiments on classification tasks show that Spectral VMamba outperforms the leading SSM models in vision, such as VMamba, while maintaining invariance to rotations and a providing a similar runtime efficiency.
****

iSegMan: Interactive Segment-and-Manipulate 3D Gaussians
Yian Zhao, Wanshi Xu, Ruochong Zheng, Pengchong Qiao, Chang Liu, Jie Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 661-670
The efficient rendering and explicit nature of 3DGS promote the advancement of 3D scene manipulation.However, existing methods typically encounter challenges in controlling the manipulation region and are unable to furnish the user with interactive feedback, which inevitably leads to unexpected results.Intuitively, incorporating interactive 3D segmentation tools can compensate for this deficiency. Nevertheless, existing segmentation frameworks impose a pre-processing step of scene-specific parameter training, which limits the efficiency and flexibility of scene manipulation.To deliver a 3D region control module that is well-suited for scene manipulation with reliable efficiency, we propose **i**nteractive **Segment-and-**Man**ipulate 3D Gaussians (**iSegMan**), an interactive segmentation and manipulation framework that only requires simple 2D user interactions in any view.To propagate user interactions to other views, we propose Epipolar-guided Interaction Propagation (**EIP**), which innovatively exploits epipolar constraint for efficient and robust interaction matching.To avoid scene-specific training to maintain efficiency, we further propose the novel Visibility-based Gaussian Voting (**VGV**), which obtains 2D segmentations from SAM and models the region extraction as a voting game between 2D Pixels and 3D Gaussians based on Gaussian visibility.Taking advantage of the efficient and precise region control of EIP and VGV, we put forth a **Manipulation Toolbox** to implement various functions on selected regions, enhancing the controllability, flexibility and practicality of scene manipulation.Extensive results on 3D scene manipulation and segmentation tasks fully demonstrate the significant advantages of iSegMan.
****

BlueLM-V-3B: Algorithm and System Co-Design for Multimodal Large Language Models on Mobile Devices
Xudong Lu, Yinghao Chen, Cheng Chen, Hui Tan, Boheng Chen, Yina Xie, Rui Hu, Guanxin Tan, Renshou Wu, Yan Hu, Yi Zeng, Lei Wu, Liuyang Bian, Zhaoxiong Wang, Long Liu, Yanzhou Yang, Han Xiao, Aojun Zhou, Yafei Wen, Xiaoxin Chen, Shuai Ren, Hongsheng Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4145-4155
The emergence and growing popularity of multimodal large language models (MLLMs) have significant potential to enhance various aspects of daily life, from improving communication to facilitating learning and problem-solving. Mobile phones, as essential daily companions, represent the most effective and accessible deployment platform for MLLMs, enabling seamless integration into everyday tasks. However, deploying MLLMs on mobile phones presents challenges due to limitations in memory size and computational capability, making it difficult to achieve smooth and real-time processing without extensive optimization. In this paper, we present BlueLM-V-3B, an algorithm and system co-design approach specifically tailored for the efficient deployment of MLLMs on mobile platforms. To be specific, we redesign the dynamic resolution scheme adopted by mainstream MLLMs and implement system optimization for hardware-aware deployment to optimize model inference on mobile phones. BlueLM-V-3B boasts the following key highlights: (1) Small Size: BlueLM-V-3B features a language model with 2.7B parameters and a vision encoder with 400M parameters. (2) Fast Speed: BlueLM-V-3B achieves a generation speed of 24.4 token/s on the MediaTek Dimensity 9300 processor with 4-bit LLM weight q

uantization. (3) Strong Performance: BlueLM-V-3B has attained the highest avera ge score of 66.1 on the OpenCompass benchmark among models with <= 4B parameters and surpassed a series of models with much larger parameter sizes (e.g., MiniCP M-V-2.6, InternVL2-8B).
****

Unraveling Normal Anatomy via Fluid-Driven Anomaly Randomization
Peirong Liu, Ana Lawry Aguila, Juan E. Iglesias; Proceedings of the Computer Vis ion and Pattern Recognition Conference (CVPR), 2025, pp. 10455-10465
Data-driven machine learning has made significant strides in medical image analy sis. However, most existing methods are tailored to specific modalities and assu me a particular resolution (often isotropic). This limits their generalizability in clinical settings, where variations in scan appearance arise from difference s in sequence parameters, resolution, and orientation. Furthermore, most general -purpose models are designed for healthy subjects and suffer from performance de gradation when pathology is present. We introduce UNA (Unraveling Normal Anatomy), the first modality-agnostic learning approach for normal brain anatomy recons truction that can handle both healthy scans and cases with pathology. We propose a fluid-driven anomaly randomization method that generates an unlimited number of realistic pathology profiles on-the-fly. UNA is trained on a combination of s ynthetic and real data, and can be applied directly to real images with potentia l pathology without the need for fine-tuning. We demonstrate UNA's effectiveness in reconstructing healthy brain anatomy and showcase its direct application to anomaly detection, using both simulated and real images from 3D healthy and stro ke datasets, including CT and MRI scans. By bridging the gap between healthy and diseased images, UNA enables the use of general-purpose models on diseased imag es, opening up new opportunities for large-scale analysis of uncurated clinical images in the presence of pathology.
****

Taming Teacher Forcing for Masked Autoregressive Video Generation
Deyu Zhou, Quan Sun, Yuang Peng, Kun Yan, Runpei Dong, Duomin Wang, Zheng Ge, Na n Duan, Xiangyu Zhang; Proceedings of the Computer Vision and Pattern Recognitio n Conference (CVPR), 2025, pp. 7374-7384
We introduce MAGI, a hybrid video generation framework that combines masked mode ling for intra-frame generation with causal modeling for next-frame generation. Our key innovation, Complete Teacher Forcing (CTF), conditions masked frames on complete observation frames rather than masked ones (namely Masked Teacher Forci ng, MTF), enabling a smooth transition from token-level (patch-level) to frame-l evel autoregressive generation. CTF significantly outperforms MTF, achieving a 2 3% improvement in FVD scores on first-frame conditioned video prediction. To add ress issues like exposure bias, we employ targeted training strategies, setting a new benchmark in autoregressive video generation. Experiments show that MAGI c an generate long, coherent video sequences exceeding 100 frames, even when train ed on as few as 16 frames, highlighting its potential for scalable, high-quality video generation.
****

UniRestore: Unified Perceptual and Task-Oriented Image Restoration Model Using D iffusion Prior
I-Hsiang Chen, Wei-Ting Chen, Yu-Wei Liu, Yuan-Chun Chiang, Sy-Yen Kuo, Ming-Hsu an Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17969-17979
Image restoration aims to recover content from inputs degraded by various factor s, such as adverse weather, blur, and noise. Perceptual Image Restoration (PIR) methods improve visual quality but often do not support downstream tasks effecti vely. On the other hand, Task-oriented Image Restoration (TIR) methods focus on enhancing image utility for high-level vision tasks, sometimes compromising visu al quality. This paper introduces UniRestore, a unified image restoration model that bridges the gap between PIR and TIR by using a diffusion prior. The diffusi on prior is designed to generate images that align with human visual quality pre ferences, but these images are often unsuitable for TIR scenarios. To solve this limitation, UniRestore utilizes encoder features from an autoencoder to adapt t

he diffusion prior to specific tasks. We propose a Complementary Feature Restoration Module (CFRM) to reconstruct degraded encoder features and a Task Feature Adapter (TFA) module to facilitate adaptive feature fusion in the decoder. This design allows UniRestore to optimize images for both human perception and downstream task requirements, addressing discrepancies between visual quality and functional needs. Integrating these modules also enhances UniRestore's adapability and efficiency across diverse tasks. Extensive expertments demonstrate the superior performance of UniRestore in both PIR and TIR scenarios.
****

Sharp-It: A Multi-view to Multi-view Diffusion Model for 3D Synthesis and Manipulation

Yiftach Edelstein, Or Patashnik, Dana Cohen-Bar, Lihi Zelnik-Manor; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21458-21468

Advancements in text-to-image diffusion models have led to significant progress in fast 3D content creation. One common approach is to generate a set of multi-view images of an object, and then reconstruct it into a 3D model. However, this approach bypasses the use of a native 3D representation of the object and is hence prone to geometric artifacts and limited in controllability and manipulation capabilities. An alternative approach involves native 3D generative models that directly produce 3D representations. These models, however, are typically limited in their resolution, resulting in lower quality 3D objects. In this work, we bridge the quality gap between methods that directly generate 3D representations and ones that reconstruct 3D objects from multi-view images. We introduce a multi-view to multi-view diffusion model called Sharp-It, which takes a 3D consistent set of multi-view images rendered from a low-quality object and enriches its geometric details and texture. The diffusion model operates on the multi-view set in parallel, in the sense that it shares features across the generated views. A high-quality 3D model can then be reconstructed from the enriched multi-view set. By leveraging the advantages of both 2D and 3D approaches, our method offers an efficient and controllable method for high-quality 3D content creation. We demonstrate that Sharp-It enables various 3D applications, such as fast synthesis, editing, and controlled generation, while attaining high-quality assets.
****

URWKV: Unified RWKV Model with Multi-state Perspective for Low-light Image Restoration

Rui Xu, Yuzhen Niu, Yuezhou Li, Huangbiao Xu, Wenxi Liu, Yuzhong Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21267-21276

Existing low-light image enhancement (LLIE) and joint LLIE and deblurring (LLIE-deblur) models have made strides in addressing predefined degradations, yet they are often constrained by dynamically coupled degradations. To address these challenges, we introduce a Unified Receptance Weighted Key Value (URWKV) model with multi-state perspective, enabling flexible and effective degradation restoration for low-light images. Specifically, we customize the core URWKV block to perceive and analyze complex degradations by leveraging multiple intra- and inter-stage states. First, inspired by the pupil mechanism in the human visual system, we propose Luminance-adaptive Normalization (LAN) that adjusts normalization parameters based on rich inter-stage states, allowing for adaptive, scene-aware luminance modulation. Second, we aggregate multiple intra-stage states through exponential moving average approach, effectively capturing subtle variations while mitigating information loss inherent in the single-state mechanism. To reduce the degradation effects commonly associated with conventional skip connections, we propose the State-aware Selective Fusion (SSF) module, which dynamically aligns and integrates multi-state features across encoder stages, selectively fusing contextual information. In comparison to state-of-the-art models, our URWKV model achieves superior performance on various benchmarks, while requiring significantly fewer parameters and computational resources. Code is available at: https://github.com/FZU-N/URWKV.
****

Revisiting Backdoor Attacks against Large Vision-Language Models from Domain Shift

Siyuan Liang, Jiawei Liang, Tianyu Pang, Chao Du, Aishan Liu, Mingli Zhu, Xiaochun Cao, Dacheng Tao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9477-9486

Instruction tuning enhances large vision-language models (LVLMs) but increases their vulnerability to backdoor attacks due to their open design. Unlike prior studies in static settings, this paper explores backdoor attacks in LVLM instruction tuning across mismatched training and testing domains. We introduce a new evaluation dimension, backdoor domain generalization, to assess attack robustness under visual and text domain shifts. Our findings reveal two insights: (1) backdoor generalizability improves when distinctive trigger patterns are independent of specific data domains or model architectures, and (2) the competitive interaction between trigger patterns and clean semantic regions, where guiding the model to predict triggers enhances attack generalizability. Based on these insights, we propose a multimodal attribution backdoor attack (MABA) that injects domain-agnostic triggers into critical areas using attributional interpretation. Experiments with OpenFlamingo, Blip-2, and Otter show that MABA significantly boosts the attack success rate of generalization by 36.4% over the unimodal attack, achieving a 97% success rate at a 0.2% poisoning rate. This study reveals limitations in current evaluations and highlights how enhanced backdoor generalizability poses a security threat to LVLMs, even without test data access.

****

Condensing Action Segmentation Datasets via Generative Network Inversion

Guodong Ding, Rongyu Chen, Angela Yao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17733-17742

This work presents the first condensation approach for procedural video datasets used in temporal action segmentation. We propose a condensation framework that leverages generative prior learned from the dataset and network inversion to condense data into compact latent codes with significant storage reduced across temporal and channel aspects. Orthogonally, we propose sampling diverse and representative action sequences to minimize video-wise redundancy. Our evaluation on standard benchmarks demonstrates consistent effectiveness in condensing TAS datasets and achieving competitive performances. Specifically, on the Breakfast dataset, our approach reduces storage by over 500xwhile retaining 83% of the performance compared to training with the full dataset. Furthermore, when applied to a downstream incremental learning task, it yields superior performance compared to the state-of-the-art.

****

TCFG: Tangential Damping Classifier-free Guidance

Mingi Kwon, Shin seong Kim, Jaeseok Jeong, Yi Ting Hsiao, Youngjung Uh; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2620-2629

Diffusion models have achieved remarkable success in text-to-image synthesis, largely attributed to the use of classifier-free guidance (CFG), which enables high-quality, condition-aligned image generation. CFG combines the conditional score (e.g., text-conditioned) with the unconditional score to control the output. However, the unconditional score is in charge of estimating the transition between manifolds of adjacent timesteps from x_t to x_{t-1}, which may inadvertently interfere with the trajectory toward the specific condition. In this work, we introduce a novel approach that leverages a geometric perspective on the unconditional score to enhance CFG performance when conditional scores are available. Specifically, we propose a method that filters the singular vectors of both conditional and unconditional scores using singular value decomposition. This filtering process aligns the unconditional score with the conditional score, thereby refining the sampling trajectory to stay closer to the manifold. Our approach improves image quality with negligible additional computation. We provide deeper insights into the score function behavior in diffusion models and present a practical technique for achieving more accurate and contextually coherent image synthesis.

****
MatAnyone: Stable Video Matting with Consistent Memory Propagation
Peiqing Yang, Shangchen Zhou, Jixin Zhao, Qingyi Tao, Chen Change Loy; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7299-7308
Auxiliary-free human video matting methods, which rely solely on input frames, often struggle with complex or ambiguous backgrounds. To tackle this, we propose MatAnyone, a practical framework designed for target-assigned video matting. Specifically, building on a memory-based framework, we introduce a consistent memory propagation module via region-adaptive memory fusion, which adaptively combines memory from the previous frame. This ensures stable semantic consistency in core regions while maintaining fine details along object boundaries. For robust training, we present a larger, high-quality, and diverse dataset for video matting. Additionally, we incorporate a novel training strategy that efficiently leverages large-scale segmentation data, further improving matting stability. With this new network design, dataset, and training strategy, MatAnyone delivers robust, accurate video matting in diverse real-world scenarios, outperforming existing methods. The code and model will be publicly available.
****
Can Generative Video Models Help Pose Estimation?
Ruojin Cai, Jason Y. Zhang, Philipp Henzler, Zhengqi Li, Noah Snavely, Ricardo Martin-Brualla; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16764-16773
Pairwise pose estimation from images with little or no overlap is an open challenge in computer vision. Existing methods, even those trained on large-scale datasets, struggle in these scenarios due to the lack of identifiable correspondences or visual overlap. Inspired by the human ability to infer spatial relationships from diverse scenes, we propose a novel approach, InterPose, that leverages the rich priors encoded within pre-trained generative video models. We propose to use a video model to hallucinate intermediate frames between two input images, effectively creating a dense, visual transition, which significantly simplifies the problem of pose estimation. Since current video models can still produce implausible motion or inconsistent geometry, we introduce a self-consistency score that evaluates the consistency of pose predictions from sampled videos. We demonstrate that our approach generalizes among three state-of-the-art video models and show consistent improvements over the state-of-the-art DUSt3R baseline on four diverse datasets encompassing indoor, outdoor, and object-centric scenes. Our findings suggest a promising avenue for improving pose estimation models by leveraging large generative models trained on vast amounts of video data, which is more readily available than 3D data. See our project page for results: Inter-Pose.github.io.
****
Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models
Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, Aniruddha Kembhavi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 91-104
Today's most advanced vision-language models (VLMs) remain proprietary. The strongest open-weight models rely heavily on synthetic data from proprietary VLMs to achieve good performance, effectively distilling these closed VLMs into open ones. As a result, the community has been missing foundational knowledge about how to build performant VLMs from scratch. We present Molmo, a new family of VLMs t

hat are state-of-the-art in their class of openness. Our key contribution is a c
ollection of new datasets called PixMo, including a dataset of highly detailed i
mage captions for pre-training, a free-form image Q&A dataset for fine-tuning, a
nd an innovative 2D pointing dataset, all collected without the use of external
VLMs. The success of our approach relies on careful modeling choices, a well-tun
ed training pipeline, and, most critically, the quality of our newly collected d
atasets. Our best-in-class 72B model not only outperforms others in the class of
 open weight and data models, but also outperforms larger proprietary models inc
luding Claude 3.5 Sonnet, and Gemini 1.5 Pro and Flash, second only to GPT-4o ba
sed on both academic benchmarks and on a large human evaluation. Our model weigh
ts, new datasets, and source code are available at https://molmo.allenai.org/blo
g.
****

DriveGPT4-V2: Harnessing Large Language Model Capabilities for Enhanced Closed-L
oop Autonomous Driving
Zhenhua Xu, Yan Bai, Yujia Zhang, Zhuoling Li, Fei Xia, Kwan-Yee K. Wong, Jianqi
ang Wang, Hengshuang Zhao; Proceedings of the Computer Vision and Pattern Recogn
ition Conference (CVPR), 2025, pp. 17261-17270
Multimodal large language models (MLLMs) possess the ability to comprehend visua
l images or videos, and show impressive reasoning ability thanks to the vast amo
unts of pretrained knowledge, making them highly suitable for autonomous driving
 applications. Unlike the previous work, DriveGPT4-V1, which focused on open-loo
p tasks, this study explores the capabilities of LLMs in enhancing closed-loop a
utonomous driving. DriveGPT4-V2 processes camera images and vehicle states as in
put to generate low-level control signals for end-to-end vehicle operation. A mu
lti-view visual tokenizer (MV-VT) is employed enabling DriveGPT4-V2 to perceive
the environment with an extensive range while maintaining critical details. The
model architecture has been refined to improve decision prediction and inference
 speed. To further enhance the performance, an additional expert LLM is trained
for online imitation learning. The expert LLM, sharing a similar structure with
DriveGPT4-V2, can access privileged information about surrounding objects for mo
re robust and reliable predictions. Experimental results show that DriveGPT4-V2
outperforms all baselines on the challenging CARLA Longest6 benchmark. The code
and data of DriveGPT4-V2 will be publicly available.
****

High-Fidelity Lightweight Mesh Reconstruction from Point Clouds
Chen Zhang, Wentao Wang, Ximeng Li, Xinyao Liao, Wanjuan Su, Wenbing Tao; Procee
dings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp
. 11739-11748
Recently, learning signed distance functions (SDFs) from point clouds has become
 popular for reconstruction. To ensure accuracy, most methods require using high
-resolution Marching Cubes for surface extraction. However, this results in redu
ndant mesh elements, making the mesh inconvenient to use. To solve the problem,
we propose an adaptive meshing method to extract resolution-adaptive meshes base
d on surface curvature, enabling the recovery of high-fidelity lightweight meshe
s. Specifically, we first use point-based representation to perceive implicit su
rfaces and calculate surface curvature. A vertex generator is designed to produc
e curvature-adaptive vertices with any specified number on the implicit surface,
 preserving the overall structure and high-curvature features. Then we develop a
 Delaunay meshing algorithm to generate meshes from vertices, ensuring geometric
 fidelity and correct topology. In addition, to obtain accurate SDFs for adaptiv
e meshing and achieve better lightweight reconstruction, we design a hybrid repr
esentation combining feature grid and feature tri-plane for better detail captur
e. Experiments demonstrate that our method can generate high-quality lightweight
 meshes from point clouds. Compared with methods from various categories, our ap
proach achieves superior results, especially in capturing more details with fewe
r elements.
****

MDP: Multidimensional Vision Model Pruning with Latency Constraint
Xinglong Sun, Barath Lakshmanan, Maying Shen, Shiyi Lan, Jingde Chen, Jose M. Al

varez; Proceedings of the Computer Vision and Pattern Recognition Conference (CV
PR), 2025, pp. 20113-20123
Current structural pruning methods face two significant limitations: (i) they of
ten limit pruning to finer-grained levels like channels, making aggressive param
eter reduction challenging, and (ii) they focus heavily on parameter and FLOP re
duction, with existing latency-aware methods frequently relying on simplistic, s
uboptimal linear models that fail to generalize well to transformers, where mult
iple interacting dimensions impact latency. In this paper, we address both limit
ations by introducing Multi-Dimensional Pruning(MDP), a novel paradigm that join
tly optimizes across a variety of pruning granularities--including channels, que
ry/key, heads, embeddings, and blocks. MDP employs an advanced latency modeling
technique to accurately capture latency variations across all prunable dimension
s, achieving an optimal balance between latency and accuracy. By reformulating p
runing as a Mixed-Integer Nonlinear Program (MINLP), MDP efficiently identifies
the optimal pruned structure across all prunable dimensions while respecting lat
ency constraints. This versatile framework supports both CNNs and transformers.
Extensive experiments demonstrate that MDP significantly outperforms previous me
thods, especially at high pruning ratios. On ImageNet, MDP achieves a 28% speed
increase with a +1.4 Top-1 accuracy improvement over prior work like HALP for Re
sNet50 pruning. Against the latest transformer pruning method, Isomorphic, MDP d
elivers an additional 37% acceleration with a +0.7 Top-1 accuracy improvement.
****

OSDFace: One-Step Diffusion Model for Face Restoration
Jingkai Wang, Jue Gong, Lin Zhang, Zheng Chen, Xing Liu, Hong Gu, Yutong Liu, Yu
lun Zhang, Xiaokang Yang; Proceedings of the Computer Vision and Pattern Recogni
tion Conference (CVPR), 2025, pp. 12626-12636
Diffusion models have demonstrated impressive performance in face restoration. Y
et, their multi-step inference process remains computationally intensive, limiti
ng their applicability in real-world scenarios. Moreover, existing methods often
 struggle to generate face images that are harmonious, realistic, and consistent
 with the subject's identity. In this work, we propose OSDFace, a novel one-step
 diffusion model for face restoration. Specifically, we propose a visual represe
ntation embedder (VRE) to better capture prior information and understand the in
put face. In VRE, low-quality faces are processed by a visual tokenizer and subs
equently embedded with a vector-quantized dictionary to generate visual prompts.
 Additionally, we incorporate a facial identity loss derived from face recogniti
on to further ensure identity consistency. We further employ a generative advers
arial network (GAN) as a guidance model to encourage distribution alignment betw
een the restored face and the ground truth. Experimental results demonstrate tha
t OSDFace surpasses current state-of-the-art (SOTA) methods in both visual quali
ty and quantitative metrics, generating high-fidelity, natural face images with
high identity consistency. The code and model will be released at https://github
.com/jkwang28/OSDFace.
****

Task Singular Vectors: Reducing Task Interference in Model Merging
Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scarda
pane, Fabrizio Silvestri, Emanuele Rodolà; Proceedings of the Computer Vision an
d Pattern Recognition Conference (CVPR), 2025, pp. 18695-18705
Task Arithmetic has emerged as a simple yet effective method to merge models wit
hout additional training. However, by treating entire networks as flat parameter
 vectors, it overlooks key structural information and is susceptible to task int
erference. In this paper, we study task vectors at the layer level, focusing on
task layer matrices and their singular value decomposition. In particular, we co
ncentrate on the resulting singular vectors, which we refer to as Task Singular
Vectors (TSV). Recognizing that layer task matrices are often low-rank, we propo
se TSV-Compress, a simple procedure that compresses them to 10% of their origina
l size while retaining 99% of accuracy. We further leverage this low-rank space
to define a new measure of task interference based on the interaction of singula
r vectors from different tasks. Building on these findings, we introduce TSV-Mer
ge, a novel model merging approach that combines compression with interference r

eduction, significantly outperforming existing methods.
****
Functionality Understanding and Segmentation in 3D Scenes
Jaime Corsetti, Francesco Giuliari, Alice Fasoli, Davide Boscaini, Fabio Poiesi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24550-24559
Understanding functionalities in 3D scenes involves interpreting natural language descriptions to locate functional interactive objects, such as handles and buttons, in a 3D environment. Functionality understanding is highly challenging, as it requires both world knowledge to interpret language and spatial perception to identify fine-grained objects. For example, given a task like "turn on the ceiling light", an embodied AI agent must infer that it needs to locate the light switch, even though the switch is not explicitly mentioned in the task description.To date, no dedicated methods have been developed for this problem. In this paper, we introduce Fun3DU, the first approach designed for functionality understanding in 3D scenes. Fun3DU uses a language model to parse the task description through Chain-of-Thought reasoning in order to identify the object of interest. The identified object is segmented across multiple views of the captured scene by using a VLM. The segmentation results from each view are lifted in 3D and aggregated into the point cloud using geometric information. Fun3DU is training-free, relying entirely on pre-trained models. We evaluate Fun3DU on SceneFun3D, the most recent and only dataset to benchmark this task, which comprises over 3000 task descriptions on 230 scenes. Our method significantly outperforms state-of-the-art open-vocabulary 3D segmentation approaches. Project page: https://tev-fbk.github.io/fun3du
****
Dragin3D: Image Editing by Dragging in 3D Space
Weiran Guang, Xiaoguang Gu, Mengqi Huang, Zhendong Mao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21502-21512
Interactive drag editing of images is a valuable task that has gained considerable attention for its precision and controllability. However, existing approaches have primarily focused on manipulating the shape or movement of objects in 2D plane. We propose to extend this drag-based editing task to 3D space. Firstly, we utilize the trajectory of two points to represent the rotational trajectory of the object. Gaussian maps of a circle and a square are centered at these two points, respectively. We use distinct shapes to ensure that symmetric views produce different object representations. Secondly, we introduce a lightweight mapping network to embed the object features into two Gaussian maps to obtain a continuous control condition that guides the model in learning the correspondence between the trajectory and the object. Finally, to overcome the limitations of current 3D object reconstruction datasets, which typically consist of object maps with transparent backgrounds, we affix random backgrounds to them. This modification helps improve the model's ability to ignore background interference when editing real images with complex backgrounds. Experiments demonstrate that our approach successfully achieves object rotation within the drag framework and demonstrates strong generalization to real-world images.
****
MMTL-UniAD: A Unified Framework for Multimodal and Multi-Task Learning in Assistive Driving Perception
Wenzhuo Liu, Wenshuo Wang, Yicheng Qiao, Qiannan Guo, Jiayin Zhu, Pengfei Li, Zilong Chen, Huiming Yang, Zhiwei Li, Lening Wang, Tiao Tan, Huaping Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6864-6874
Advanced driver assistance systems require a comprehensive understanding of the driver's mental/physical state and traffic context but existing works often neglect the potential benefits of joint learning between these tasks. This paper proposes MMTL-UniAD, a unified multi-modal multi-task learning framework that simultaneously recognizes driver behavior (e.g., looking around, talking), driver emotion (e.g., anxiety, happiness), vehicle behavior (e.g., parking, turning), and traffic context (e.g., traffic jam, traffic smooth). A key challenge is avoiding

negative transfer between tasks, which can impair learning performance. To addr
ess this, we introduce two key components into the framework: one is the multi-a
xis region attention network to extract global context-sensitive features, and t
he other is the dual-branch multimodal embedding to learn multimodal embeddings
from both task-shared and task-specific features. The former uses a multi-attent
ion mechanism to extract task-relevant features, mitigating negative transfer ca
used by task-unrelated features. The latter employs a dual-branch structure to a
daptively adjust task-shared and task-specific parameters, enhancing cross-task
knowledge transfer while reducing task conflicts. We assess MMTL-UniAD on the AI
DE dataset, using a series of ablation studies, and show that it outperforms sta
te-of-the-art methods across all four tasks. The code is available on https://gi
thub.com/Wenzhuo-Liu/MMTL-UniAD.
****

T2V-CompBench: A Comprehensive Benchmark for Compositional Text-to-video Generat
ion
Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, Xihui Liu; Proc
eedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025,
pp. 8406-8416
Text-to-video (T2V) generative models have advanced significantly, yet their abi
lity to compose different objects, attributes, actions, and motions into a video
remains unexplored. Previous text-to-video benchmarks also neglect this importa
nt ability for evaluation. In this work, we conduct the first systematic study o
n compositional text-to-video generation. We propose T2V-CompBench, the first be
nchmark tailored for compositional text-to-video generation. T2V-CompBench encom
passes diverse aspects of compositionality, including consistent attribute bindi
ng, dynamic attribute binding, spatial relationships, motion binding, action bin
ding, object interactions, and generative numeracy. We further carefully design
evaluation metrics of multimodal large language model (MLLM)-based, detection-ba
sed, and tracking-based metrics, which can better reflect the compositional text
-to-video generation quality of seven proposed categories with 1400 text prompts
. The effectiveness of the proposed metrics is verified by correlation with huma
n evaluations. We also benchmark various text-to-video generative models and con
duct in-depth analysis across different models and various compositional categor
ies. We find that compositional text-to-video generation is highly challenging f
or current models, and we hope our attempt could shed light on future research i
n this direction.
****

Self-Evolving Visual Concept Library using Vision-Language Critics
Atharva Sehgal, Patrick Yuan, Ziniu Hu, Yisong Yue, Jennifer J. Sun, Swarat Chau
dhuri; Proceedings of the Computer Vision and Pattern Recognition Conference (CV
PR), 2025, pp. 13124-13134
We study the problem of building a visual concept library for visual recognition
. Building effective visual concept libraries is challenging, as manual definiti
on is labor-intensive, while relying solely on LLMs for concept generation can r
esult in concepts that lack discriminative power or fail to account for the comp
lex interactions between them. Our approach, ESCHER, takes a library learning pe
rspective to iteratively discover and improve visual concepts. ESCHER uses a vis
ion-language model (VLM) as a critic to iteratively refine the concept library,
including accounting for interactions between concepts and how they affect downs
tream classifiers. By leveraging the in-context learning abilities of LLMs and t
he history of performance using various concepts, ESCHER dynamically improves it
s concept generation strategy based on the VLM critic's feedback. Finally, ESCHE
R does not require any human annotations, and is thus an automated plug-and-play
framework. We empirically demonstrate the ability of ESCHER to learn a concept
library for zero-shot, few-shot, and fine-tuning visual classification tasks. Th
is work represents, to our knowledge, the first application of concept library l
earning to real-world visual tasks.
****

Multimodal Autoregressive Pre-training of Large Vision Encoders
Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Hald

imann, Sai Aitharaju, Victor G. Turrisi da Costa, Louis Béthune, Zhe Gan, Alexander Toshev, Marcin Eichner, Moin Nabi, Yinfei Yang, Joshua Susskind, Alaaeldin El-Nouby; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9641-9654
We introduce a novel method for pre-training of large-scale vision encoders. Building on recent advancements in autoregressive pre-training of vision models, we extend this framework to a multimodal setting, i.e., images and text. In this paper, we present AIMV2, a family of generalist vision encoders characterized by a straightforward pre-training process, scalability, and remarkable performance across a range of downstream tasks. This is achieved by pairing the vision encoder with a multimodal decoder that autoregressively generates raw image patches and text tokens. Our encoders excel not only in multimodal evaluations but also in vision benchmarks such as localization, grounding, and classification. Notably, our AIMV2-3B encoder achieves 89.5% accuracy on ImageNet-1k with a frozen trunk. Fur- thermore, AIMV2 consistently outperforms state-of-the-art contrastive models (e.g., CLIP, SigLIP) in multimodal im- age understanding across diverse settings.
****

AKiRa: Augmentation Kit on Rays for Optical Video Generation
Xi Wang, Robin Courant, Marc Christie, Vicky Kalogeiton; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2609-2619
Recent advances in text-conditioned video diffusion have greatly improved video quality. However, these methods offer limited or sometimes no control to users on camera aspects, including dynamic camera motion, zoom, distorted lens and focus shifts. These motion and optical aspects are crucial for adding controllability and cinematic elements to generation frameworks, ultimately resulting in visual content that draws focus, enhances mood, and guides emotions according to film makers' controls. In this paper, we aim to close the gap between controllable video generation and camera optics. To achieve this, we propose AKiRa (Augmentation Kit on Rays), a novel augmentation framework that builds and trains a camera adapter with a complex camera model over an existing video generation backbone. It enables fine-tuned control over camera motion as well as complex optical parameters (focal length, distortion, aperture) to achieve cinematic effects such as zoom, fisheye effect, and bokeh. Extensive experiments demonstrate AKiRa's effectiveness in combining and composing camera optics while outperforming all state-of-the-art methods. This work sets a new landmark in controlled and optically enhanced video generation, paving the way for future camera diffusion methods.
****

Towards Stable and Storage-efficient Dataset Distillation: Matching Convexified Trajectory
Wenliang Zhong, Haoyu Tang, Qinghai Zheng, Mingzhu Xu, Yupeng Hu, Weili Guan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25581-25589
The rapid evolution of deep learning and large language models has led to an exponential growth in the demand for training data, prompting the development of Dataset Distillation methods to address the challenges of managing large datasets. Among these, Matching Training Trajectories (MTT) has been a prominent approach, which replicates the training trajectory of an expert network on real data with a synthetic dataset. However, our investigation found that this method suffers from three significant limitations: 1. Instability of expert trajectory generated by Stochastic Gradient Descent (SGD); 2. Low convergence speed of the distillation process; 3. High storage consumption of the expert trajectory. To address these issues, we offer a new perspective on understanding the essence of Dataset Distillation and MTT through a simple transformation of the objective function, and introduce a novel method called Matching Convexified Trajectory (MCT), which aims to provide better guidance for the student trajectory. MCT creates convex combinations of expert trajectories by selecting a few expert models, guiding student networks to converge quickly and stably. This trajectory is not only easier to store, but also enables continuous sampling strategies during the distillation process, ensuring thorough learning and fitting of the entire expert trajec

tory. The comprehensive experiment of three public datasets verified that MCT is superior to the traditional MTT method.
****

TSAM: Temporal SAM Augmented with Multimodal Prompts for Referring Audio-Visual Segmentation
Abduljalil Radman, Jorma Laaksonen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23947-23956
Referring audio-visual segmentation (Ref-AVS) aims to segment objects within audio-visual scenes using multimodal cues embedded in text expressions. While the Segment Anything Model (SAM) has revolutionized visual segmentation, its applicability to Ref-AVS, where multimodal cues act as novel prompts, remains unexplored. SAM's limitation to single-frame segmentation also hinders its ability to capture essential temporal context needed for multi-frame audio-visual segmentation. To address this gap, we propose TSAM, a novel extension of SAM designed to leverage multimodal cues for precise segmentation in dynamic audio-visual scenes. TSAM enhances SAM's image encoder with a temporal modeling branch, enabling spatio-temporal learning and deep multimodal fusion across video frames, while retaining SAM's pre-trained knowledge. Additionally, TSAM replaces SAM's user-interactive prompting mechanism with sparse and dense data-driven prompts, enabling more effective integration of audio-visual inputs and reference text expressions. Extensive experiments on the Ref-AVS dataset demonstrate TSAM's superiority over state-of-the-art methods. The results illustrate its effectiveness in segmenting objects in dynamic audio-visual scenes using text-based multimodal cues and its strong generalization to unseen objects.
****

TFCustom: Customized Image Generation with Time-Aware Frequency Feature Guidance
Mushui Liu, Dong She, Jingxuan Pang, Qihan Huang, Jiacheng Ying, Wanggui He, Yuanlei Hou, Siming Fu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2714-2723
Subject-driven image personalization has seen notable advancements, especially with the advent of the ReferenceNet paradigm. ReferenceNet excels in integrating image reference features, making it highly applicable in creative and commercial settings. However, current implementations of ReferenceNet primarily operate as latent-level feature extractors, which limit their potential. This constraint hinders the provision of appropriate features to the denoising backbone across different timesteps, leading to suboptimal image consistency. In this paper, we revisit the extraction of reference features and propose TFCustom, a model framework designed to focus on reference image features at different temporal steps and frequency levels. Specifically, we firstly propose synchronized ReferenceNet to extract reference image features while simultaneously optimizing noise injection and denoising for the reference image. We also propose a time-aware frequency feature refinement module that leverages high- and low-frequency filters, combined with time embeddings, to adaptively select the degree of reference feature injection. Additionally, to enhance the similarity between reference objects and the generated image, we introduce a novel reward-based loss that encourages greater alignment between the reference and generated images. Experimental results demonstrate state-of-the-art performance in both multi-object and single-object reference generation, with significant improvements in texture and textual detail generation over existing methods.
****

Boosting Point-Supervised Temporal Action Localization through Integrating Query Reformation and Optimal Transport
Mengnan Liu, Le Wang, Sanping Zhou, Kun Xia, Xiaolong Sun, Gang Hua; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13865-13875
Point-supervised Temporal Action Localization poses significant challenges due to the difficulty of identifying complete actions with a single-point annotation per action. Existing methods typically employ Multiple Instance Learning, which struggles to capture global temporal context and requires heuristic post-processing. In research on fully-supervised tasks, DETR-based structures have effectiv

ely addressed these limitations. However, it is nontrivial to merely adapt DETR to this task, encountering two major bottlenecks. (1) How to integrate point label information into the model and (2) How to select optimal decoder proposals for training in the absence of complete action segment annotations. To address this issue, we introduce an end-to-end framework by integrating Query Reformation and Optimal Transport (QROT). Specifically, we encode point labels through a set of semantic consensus queries, enabling effective focus on action-relevant snippets. Furthermore, we integrate an optimal transport mechanism to generate high-quality pseudo labels. These pseudo-labels facilitate precise proposals selection based on Hungarian algorithm, significantly enhancing localization accuracy in point-supervised settings. Extensive experiments on the THUMOS14 and ActivityNet-v1.3 datasets demonstrate that our method outperforms existing MIL-based approaches, offering more stable and accurate temporal action localization in point-level supervision.

****

SketchFusion: Learning Universal Sketch Features through Fusing Foundation Models

Subhadeep Koley, Tapas Kumar Dutta, Aneeshan Sain, Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Yi-Zhe Song; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2556-2567

While foundation models have revolutionised computer vision, their effectiveness for sketch understanding remains limited by the unique challenges of abstract, sparse visual inputs. Through systematic analysis, we uncover two fundamental limitations: Stable Diffusion (SD) struggles to extract meaningful features from abstract sketches (unlike its success with photos), and exhibits a pronounced frequency-domain bias that suppresses essential low-frequency components needed for sketch understanding. Rather than costly retraining, we address these limitations by strategically combining SD with CLIP, whose strong semantic understanding naturally compensates for SD's spatial-frequency biases. By dynamically injecting CLIP features into SD's denoising process and adaptively aggregating features across semantic levels, our method achieves state-of-the-art performance in sketch retrieval (+3.35%), recognition (+1.06%), segmentation (+29.42%), and correspondence learning (+21.22%), demonstrating the first truly universal sketch feature representation in the era of foundation models.

****

Bridging the Vision-Brain Gap with an Uncertainty-Aware Blur Prior

Haitao Wu, Qing Li, Changqing Zhang, Zhen He, Xiaomin Ying; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2246-2257

Can our brain signals faithfully reflect the original visual stimuli, even including high-frequency details? Although human perceptual and cognitive capacities enable us to process and remember visual information, these abilities are constrained by several factors, such as limited attentional resources and finite capacity of visual memory. When visual stimuli are processed by human visual system into brain signals, some information is inevitably lost, leading to a discrepancy known as the System GAP. Additionally, perceptual and cognitive dynamics, along with technical noise in signal acquisition, degrade the fidelity of brain signals relative to the visual stimuli, known as the Random GAP. When encoded brain representations are directly aligned with the corresponding pretrained image features, the System GAP and Random GAP between paired data challenge the model, requiring it to bridge these gaps. However, due to limited paired data, these gaps are difficult for the model to learn, leading to overfitting and poor generalization to new data. To address these GAPs, we propose a simple yet effective approach called the Uncertainty-aware Blur Prior (UBP). It estimates the uncertainty within the paired data, reflecting the mismatch between brain signals and visual stimuli. Based on uncertainty, UBP dynamically blurs the high-frequency details of the original images, reducing the impact of mismatch and improving alignment. Our method achieves a top-1 accuracy of 50.9% and a top-5 accuracy of 79.7% on the zero-shot brain-to-image retrieval task, surpassing previous state-of-the-art methods. Code is available at https://github.com/HaitaoWuTJU/Uncertainty-aware-Blur-Prior.

****
Invisible Backdoor Attack against Self-supervised Learning
Hanrong Zhang, Zhenting Wang, Boheng Li, Fulin Lin, Tingxu Han, Mingyu Jin, Chen lu Zhan, Mengnan Du, Hongwei Wang, Shiqing Ma; Proceedings of the Computer Visio n and Pattern Recognition Conference (CVPR), 2025, pp. 25790-25801
Self-supervised learning (SSL) models are vulnerable to backdoor attacks. Existi ng backdoor attacks that are effective in SSL often involve noticeable triggers, like colored patches or visible noise, which are vulnerable to human inspection . This paper proposes an imperceptible and effective backdoor attack against sel f-supervised models. We first find that existing imperceptible triggers designed for supervised learning are less effective in compromising self-supervised mode ls. We then identify this ineffectiveness is attributed to the overlap in distri butions between the backdoor and augmented samples used in SSL. Building on this insight, we design an attack using optimized triggers disentangled with the aug mented transformation in the SSL, while remaining imperceptible to human vision. Experiments on five datasets and six SSL algorithms demonstrate our attack is h ighly effective and stealthy. It also has strong resistance to existing backdoor defenses. Our code can be found at https://github.com/Zhang-Henry/INACTIVE.
****
Perceptually Accurate 3D Talking Head Generation: New Definitions, Speech-Mesh R epresentation, and Evaluation Metrics
Lee Chae-Yeon, Oh Hyun-Bin, Han EunGi, Kim Sung-Bin, Suekyeong Nam, Tae-Hyun Oh; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2 025, pp. 21065-21074
Recent advancements in speech-driven 3D talking head generation have made signif icant progress in lip synchronization. However, existing models still struggle t o capture the perceptual alignment between varying speech characteristics and co rresponding lip movements. In this work, we claim that three criteria--Temporal Synchronization, Lip Readability, and Expressiveness--are crucial for achieving perceptually accurate lip movements. Motivated by our hypothesis that a desirabl e representation space exists to meet these three criteria, we introduce a speec h-mesh synchronized representation that captures intricate correspondences betwe en speech signals and 3D face meshes. We found that our learned representation e xhibits desirable characteristics, and we plug it into existing models as a perc eptual loss to better align lip movements to the given speech. In addition, we u tilize this representation as a perceptual metric and introduce two other physic ally grounded lip synchronization metrics to assess how well the generated 3D ta lking heads align with these three criteria. Experiments show that training 3D t alking head generation models with our perceptual loss significantly improve all three aspects of perceptually accurate lip synchronization. Codes and datasets are available at https://perceptual-3d-talking-head.github.io/.
****
BWFormer: Building Wireframe Reconstruction from Airborne LiDAR Point Cloud with Transformer
Yuzhou Liu, Lingjie Zhu, Hanqiao Ye, Shangfeng Huang, Xiang Gao, Xianwei Zheng, Shuhan Shen; Proceedings of the Computer Vision and Pattern Recognition Conferen ce (CVPR), 2025, pp. 22215-22224
In this paper, we present BWFormer, a novel Transformer-based model for building wireframe reconstruction from airborne LiDAR point cloud. The problem is solve d in a ground-up manner here by detecting the building corners in 2D, lifting an d connecting them in 3D space afterwards with additional data augmentation.Due t o the 2.5D characteristic of the airborne LiDAR point cloud, we simplify the pr oblem by projecting the points on the ground plane to produce a 2D height map. W ith the height map, a heat map is first generated with pixel-wise corner likelih ood to predict the possible 2D corners.Then, 3D corners are predicted by a Trans former-based network with extra height embedding initialization. This 2D-to-3D c orner detection strategy reduces the search space significantly. To recover the topological connections among the corners, edges are finally predicted from the height map with the proposed edge attention mechanism, which extracts holistic f eatures and preserves local details simultaneously. In addition, due to the limi

ted datasets in the field and the irregularity of the point clouds, a conditiona l latent diffusion model for LiDAR scanning simulation is utilized for data augm entation. BWFormer surpasses other state-of-the-art methods, especially in recon struction completeness. Our code is available at: https://github.com/3dv-casia/ BWformer/.
****

Diffusion-4K: Ultra-High-Resolution Image Synthesis with Latent Diffusion Models
Jinjin Zhang, Qiuyu Huang, Junjie Liu, Xiefan Guo, Di Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23464-23473
In this paper, we present Diffusion-4K, a novel framework for direct ultra-high-resolution image synthesis using text-to-image diffusion models. The core advanc ements include: (1) Aesthetic-4K Benchmark: addressing the absence of a publicly available 4K image synthesis dataset, we construct Aesthetic-4K, a comprehensiv e benchmark for ultra-high-resolution image generation. We curated a high-qualit y 4K dataset with carefully selected images and captions generated by GPT-4o. Ad ditionally, we introduce GLCM Score and compression ratio metrics to evaluate fi ne details, combined with holistic measures such as FID, Aesthetics and CLIPScor e for a comprehensive assessment of ultra-high-resolution images. (2) Wavelet-ba sed Fine-tuning: we propose a wavelet-based fine-tuning approach for direct trai ning with photorealistic 4K images, applicable to various latent diffusion model s, demonstrating its effectiveness in synthesizing highly detailed 4K images. Co nsequently, Diffusion-4K achieves impressive performance in high-quality image s ynthesis and text prompt adherence, especially when powered by modern large-scal e diffusion models (e.g., SD3-2B and Flux-12B). Extensive experimental results f rom our benchmark demonstrate the superiority of Diffusion-4K in ultra-high-reso lution image synthesis. Code is available at https://github.com/zhang0jhon/diffu sion-4k.
****

AffordDP: Generalizable Diffusion Policy with Transferable Affordance
Shijie Wu, Yihang Zhu, Yunao Huang, Kaizhen Zhu, Jiayuan Gu, Jingyi Yu, Ye Shi, Jingya Wang; Proceedings of the Computer Vision and Pattern Recognition Conferen ce (CVPR), 2025, pp. 6971-6980
Diffusion-based policies have shown impressive performance in robotic manipulati on tasks while struggling with out-of-domain distributions. Recent efforts attem pted to enhance generalization by improving the visual feature encoding for diff usion policy. However, their generalization is typically limited to the same cat egory with similar appearances. Our key insight is that leveraging affordances--manipulation priors that define "where" and "how" an agent interacts with an obj ect--can substantially enhance generalization to entirely unseen object instance s and categories. We introduce the Diffusion Policy with transferable Affordance (AffordDP), designed for generalizable manipulation across novel categories. Af fordDP models affordances through 3D contact points and post-contact trajectorie s, capturing the essential static and dynamic information for complex tasks. The transferable affordance from in-domain data to unseen objects is achieved by es timating a 6D transformation matrix using foundational vision models and point c loud registration techniques. More importantly, we incorporate affordance guidan ce during diffusion sampling that can refine action sequence generation. This gu idance directs the generated action to gradually move towards the desired manipu lation for unseen objects while keeping the generated action within the manifold of action space. Experimental results from both simulated and real-world enviro nments demonstrate that AffordDP consistently outperforms previous diffusion-bas ed methods, successfully generalizing to unseen instances and categories where o thers fail.
****

HMAR: Efficient Hierarchical Masked Auto-Regressive Image Generation
Hermann Kumbong, Xian Liu, Tsung-Yi Lin, Ming-Yu Liu, Xihui Liu, Ziwei Liu, Dani el Y. Fu, Christopher Re, David W. Romero; Proceedings of the Computer Vision an d Pattern Recognition Conference (CVPR), 2025, pp. 2535-2544
Visual AutoRegressive modeling (VAR) shows promise in bridging the speed and qua lity gap between autoregressive image models and diffusion models. VAR reformula

tes autoregressive modeling by decomposing an image into successive resolution s cales. During inference, an image is generated by predicting all the tokens in t he next (higher-resolution) scale, conditioned on all tokens in all previous (lo wer-resolutions) scales. However, this formulation suffers from reduced image qu ality due to parallel generation of all tokens in a resolution scale; has sequen ce lengths scaling superlinearly in image resolution; and requires retraining to change the resolution sampling schedule.We introduce \underline H ierarchical \ underline M asked \underline A uto\underline R egressive modeling (HMAR), a new image generation algorithm that alleviates these issues using next-scale predict ion and masked prediction to generate high-quality images with fast sampling. HM AR reformulates next-scale prediction as a Markovian process, wherein prediction of each resolution scale is conditioned only on tokens in its immediate predece ssor instead of the tokens in all predecessor resolutions. When predicting a res olution scale, HMAR uses a controllable multi-step masked generation procedure t o generate a subset of the tokens in each step. On ImageNet 256 x 256 and 512 x 512 benchmarks, HMAR models match or outperform parameter-matched VAR, diffu sion, and autoregressive baselines. We develop efficient IO-aware block-sparse a ttention kernels that allow HMAR to achieve faster training and inference times over VAR by over 2.5xand 1.75 xrespectively, as well as over 3 xlower inference memory footprint. Finally, the Markovian formulation of HMAR yields additional f lexibility over VAR; we show that its sampling schedule can be changed without f urther training, and it can be applied to image editing tasks in a zero-shot man ner.

****

OmniDrive: A Holistic Vision-Language Dataset for Autonomous Driving with Counte rfactual Reasoning

Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Ka utz, Ying Li, Jose M. Alvarez; Proceedings of the Computer Vision and Pattern Re cognition Conference (CVPR), 2025, pp. 22442-22452

The advances in vision-language models (VLMs) have led to a growing interest in autonomous driving to leverage their strong reasoning capabilities. However, ext ending these capabilities from 2D to full 3D understanding is crucial for real-w orld applications. To address this challenge, we propose OmniDrive, a holistic v ision-language dataset that aligns agent models with 3D driving tasks through co unterfactual reasoning. This approach enhances decision-making by evaluating pot ential scenarios and their outcomes, similar to human drivers considering altern ative actions. Our counterfactual-based synthetic data annotation process genera tes large-scale, high-quality datasets, providing denser supervision signals tha t bridge planning trajectories and language-based reasoning. Futher, we explore two advanced OmniDrive-Agent frameworks, namely Omni-L and Omni-Q, to assess the importance of vision-language alignment versus 3D perception, revealing critica l insights into designing effective LLM-agents. Significant improvements on the DriveLM Q&A benchmark and nuScenes open-loop planning demonstrate the effectiven ess of our dataset and methods.

****

DKC: Differentiated Knowledge Consolidation for Cloth-Hybrid Lifelong Person Re-identification

Zhenyu Cui, Jiahuan Zhou, Yuxin Peng; Proceedings of the Computer Vision and Pat tern Recognition Conference (CVPR), 2025, pp. 3573-3582

Lifelong person re-identification (LReID) aims to match the same person using se quentially collected data. However, due to the long-term nature of lifelong lear ning, the inevitable changes in human clothes prevent the model from relying on unified discriminative information (e.g., clothing style) to match the same pers on in the streaming data, demanding differentiated cloth-irrelevant information. Unfortunately, existing LReID methods typically fail to leverage such knowledge resulting in the exacerbation of catastrophic forgetting issues. Therefore, in this paper, we focus on a challenging practical task called Cloth-Hybrid Lifelon g Person Re-identification (CH-LReID), which requires matching the same person w earing different clothes using sequentially collected data. A Differentiated Kno wledge Consolidation (DKC) framework is designed to unify and balance distinct k

nowledge across streaming data. The core idea is to adaptively balance different iated knowledge and compatibly consolidate cloth-relevant and cloth-irrelevant i nformation. To this end, a Differentiated Knowledge Transfer (DKT) module and a Latent Knowledge Consolidation (LKC) module are designed to adaptively discover differentiated new knowledge, while eliminating the derived domain shift of old knowledge via reconstructing the old latent feature space, respectively. Then, t o further alleviate the catastrophic conflict between differentiated new and old knowledge, we further propose a Dual-level Distribution Alignment (DDA) module to align the distribution of discriminative knowledge at both the instance level and the fine-grained level. Extensive experiments on multiple benchmarks demons trate the superiority of our method against existing methods in both CH-LReID an d traditional LReID tasks. The source code of this paper is available at https:/ /github.com/PKU-ICST-MIPL/DKC-CVPR2025.
****

Enhancing Facial Privacy Protection via Weakening Diffusion Purification
Ali Salar, Qing Liu, Yingli Tian, Guoying Zhao; Proceedings of the Computer Visi on and Pattern Recognition Conference (CVPR), 2025, pp. 8235-8244
The rapid growth of social media has led to the widespread sharing of individual portrait images, which pose serious privacy risks due to the capabilities of au tomatic face recognition (AFR) systems for mass surveillance. Hence, protecting facial privacy against unauthorized AFR systems is essential. Inspired by the ge neration capability of the emerging diffusion models, recent methods employ diff usion models to generate adversarial face images for privacy protection. However , they suffer from the diffusion purification effect, leading to a low protectio n success rate (PSR). In this paper, we first propose learning unconditional emb eddings to increase the learning capacity for adversarial modifications and then use them to guide the modification of the adversarial latent code to weaken the diffusion purification effect. Moreover, we integrate an identity-preserving st ructure to maintain structural consistency between the original and generated im ages, allowing human observers to recognize the generated image as having the sa me identity as the original. Extensive experiments conducted on two public datas ets, i.e., CelebA-HQ and LADN, demonstrate the superiority of our approach. The protected faces generated by our method outperform those produced by existing fa cial privacy protection approaches in terms of transferability and natural appea rance. The code is available at https://github.com/parham1998/Facial-Privacy-Pro tection
****

ORIDa: Object-centric Real-world Image Composition Dataset
Jinwoo Kim, Sangmin Han, Jinho Jeong, Jiwoo Choi, Dongyeoung Kim, Seon Joo Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 20 25, pp. 3051-3060
Object compositing, the task of placing and harmonizing objects in images of div erse visual scenes, has become an important task in computer vision with the ris e of generative models.However, existing datasets lack the diversity and scale r equired to comprehensively explore real-world scenarios comprehensively. We intr oduce ORIDa (Object-centric Real-world Image Composition Dataset), a large-scale , real-captured dataset containing over 30,000 images featuring 200 unique objec ts, each of which is presented across varied positions and scenes. ORIDa has two types of data: factual-counterfactual sets and factual-only scenes. The factual -counterfactual sets consist of four factual images showing an object in differe nt positions within a scene and a single counterfactual (or background) image of the scene without the object, resulting in five images per scene. The factual-o nly scenes include a single image containing an object in a specific context, ex panding the variety of environments. To our knowledge, ORIDa is the first public ly available dataset with its scale and complexity for real-world image composit ion. Extensive analysis and experiments highlight the value of ORIDa as a resour ce for advancing further research in object compositing.
****

MeGA: Hybrid Mesh-Gaussian Head Avatar for High-Fidelity Rendering and Head Edit ing

Cong Wang, Di Kang, Heyi Sun, Shenhan Qian, Zixuan Wang, Linchao Bao, Song-Hai Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26274-26284

Creating high-fidelity head avatars from multi-view videos is essential for many AR/VR applications. However, current methods often struggle to achieve high-quality renderings across all head components (e.g., skin vs. hair) due to the limitations of using one single representation for elements with varying characteristics. In this paper, we introduce a Hybrid Mesh-Gaussian Head Avatar (MeGA) that models different head components with more suitable representations. Specifically, we employ an enhanced FLAME mesh for the facial representation and predict a UV displacement map to provide per-vertex offsets for improved personalized geometric details. To achieve photorealistic rendering, we use deferred neural rendering to obtain facial colors and decompose neural textures into three meaningful parts. For hair modeling, we first build a static canonical hair using 3D Gaussian Splatting. A rigid transformation and an MLP-based deformation field are further applied to handle complex dynamic expressions. Combined with our occlusion-aware blending, MeGA generates higher-fidelity renderings for the whole head and naturally supports diverse downstream tasks. Experiments on the NeRSemble dataset validate the effectiveness of our designs, outperforming previous state-of-the-art methods and enabling versatile editing capabilities, including hairstyle alteration and texture editing.

****

Image Generation Diversity Issues and How to Tame Them

Mischa Dombrowski, Weitong Zhang, Sarah Cechnicka, Hadrien Reynaud, Bernhard Kainz; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3029-3039

Generative methods have reached a level of quality that is almost indistinguishable from real data. However, while individual samples may appear unique, generative models often exhibit limitations in covering the full data distribution. Unlike quality issues, diversity problems within generative models are not easily detected by simply observing single images or generated datasets, which means we need a specific measure to assess the diversity of these models. In this paper, we draw attention to the current lack of diversity in generative models and the inability of common metrics to measure this. We achieve this by framing diversity as an image retrieval problem, where we measure how many real images can be retrieved using synthetic data as queries. This yields the Image Retrieval Score (IRS), an interpretable, hyperparameter-free metric that quantifies the diversity of a generative model's output. IRS requires only a subset of synthetic samples and provides a statistical measure of confidence. Our experiments indicate that current feature extractors commonly used in generative model assessment are inadequate for evaluating diversity effectively. Consequently, we perform an extensive search for the best feature extractors to assess diversity. Evaluation reveals that current diffusion models converge to limited subsets of the real distribution, with no current state-of-the-art models superpassing 77% of the diversity of the training data. To address this limitation, we introduce Diversity-Aware Diffusion Models (DiADM), a novel approach that improves diversity of unconditional diffusion models without loss of image quality. We do this by disentangling diversity from image quality by using a diversity aware module that uses pseudo-unconditional features as input. We provide a Python package offering unified feature extraction and metric computation to further facilitate the evaluation of generative models https://github.com/MischaD/beyondfid.

****

Annotation Ambiguity Aware Semi-Supervised Medical Image Segmentation

Suruchi Kumari, Pravendra Singh; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10404-10413

Despite the remarkable progress of deep learning-based methods in medical image segmentation, their use in clinical practice remains limited for two main reasons. First, obtaining a large medical dataset with precise annotations to train segmentation models is challenging. Secondly, most current segmentation techniques generate a single deterministic segmentation mask for each image. However, in r

eal-world scenarios, there is often significant uncertainty regarding what defin es the "correct" segmentation, and various expert annotators might provide diffe rent segmentations for the same image. To tackle both of these problems, we prop ose Annotation Ambiguity Aware Semi-Supervised Medical Image Segmentation (AmbiS SL). AmbiSSL combines a small amount of multi-annotator labeled data and a large set of unlabeled data to generate diverse and plausible segmentation maps. Our method consists of three key components: (1) The Diverse Pseudo-Label Generation (DPG) module utilizes multiple decoders, created by performing randomized pruni ng on the original backbone decoder. These pruned decoders enable the generation of a diverse pseudo-label set; (2) a Semi-Supervised Latent Distribution Learni ng (SSLDL) module constructs a common latent space by utilizing both ground trut h annotations and pseudo-label set; and (3) a Cross-Decoder Supervision (CDS) mo dule, which enables pruned decoders to guide each other's learning. We evaluated the proposed method on two publicly available datasets. Extensive experiments d emonstrate that AmbiSSL can generate diverse segmentation maps using only a smal l amount of labeled data and abundant unlabeled data, offering a more practical solution for medical image segmentation by reducing reliance on large labeled da tasets.

****

Effective Cloud Removal for Remote Sensing Images by an Improved Mean-Reverting Denoising Model with Elucidated Design Space

Yi Liu, Wengen Li, Jihong Guan, Shuigeng Zhou, Yichao Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17851-17861

Cloud removal (CR) remains a challenging task in remote sensing image processing . Although diffusion models (DM) exhibit strong generative capabilities, their d irect applications to CR are suboptimal, as they generate cloudless images from random noise, ignoring inherent information in cloudy inputs. To overcome this d rawback, we develop a new CR model EMRDM based on mean-reverting diffusion model s (MRDMs) to establish a direct diffusion process between cloudy and cloudless i mages. Compared to current MRDMs, EMRDM offers a modular framework with updatabl e modules and an elucidated design space, based on a reformulated forward proces s and a new ordinary differential equation (ODE)-based backward process. Leverag ing our framework, we redesign key MRDM modules to boost CR performance, includi ng restructuring the denoiser via a preconditioning technique, reorganizing the training process, and improving the sampling process by introducing determinisli c and stochastic samplers. To achieve multi-temporal CR, we further develop a de noising network for simultaneously denoising sequential images. Experiments on m ono-temporal and multi-temporal datasets demonstrate the superior performance of EMRDM. Our code is available at https://github.com/Ly403/EMRDM.

****

CAP4D: Creating Animatable 4D Portrait Avatars with Morphable Multi-View Diffusi on Models

Felix Taubner, Ruihang Zhang, Mathieu Tuli, David B. Lindell; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5318-5330

Reconstructing photorealistic and dynamic portrait avatars from images is essent ial to many applications including advertising, visual effects, and virtual real ity. Depending on the application, avatar reconstruction involves different capt ure setups and constraints -- for example, visual effects studios use camera arr ays to capture hundreds of reference images, while content creators may seek to animate a single portrait image downloaded from the internet. As such, there is a large and heterogeneous ecosystem of methods for avatar reconstruction. Techni ques based on multi-view stereo or neural rendering achieve the highest quality results, but require hundreds of reference images. Recent generative models prod uce convincing avatars from a single reference image, but visual fidelity lags b ehind multi-view techniques. Here, we present CAP4D: an approach that uses a mor phable multi-view diffusion model to reconstruct photoreal 4D (dynamic 3D) portr ait avatars from any number of reference images (i.e., one to 100) and animate a nd render them in real time. Our approach demonstrates state-of-the-art performa nce for single-, few-, and multi-image 4D portrait avatar reconstruction, and ta kes steps to bridge the gap in visual fidelity between single-image and multi-vi

ew reconstruction techniques.
****
Comprehensive Information Bottleneck for Unveiling Universal Attribution to Interpret Vision Transformers
Jung-Ho Hong, Ho-Joong Kim, Kyu-Sung Jeon, Seong-Whan Lee; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25166-25175
The feature attribution method reveals the contribution of input variables to the decision-making process to provide an attribution map for explanation. Existing methods grounded on the information bottleneck principle compute information in a specific layer to obtain attributions, compressing the features by injecting noise via a parametric damping ratio. However, the attribution obtained in a specific layer neglects evidence of the decision-making process distributed across layers. In this paper, we introduce a comprehensive information bottleneck (CoIBA), which discovers the relevant information in each targeted layer to explain the decision-making process. Our core idea is applying information bottleneck in multiple targeted layers to estimate the comprehensive information by sharing a parametric damping ratio across the layers. Leveraging this shared ratio complements the over-compressed information to discover the omitted clues of the decision by sharing the relevant information across the targeted layers. We suggest the variational approach to fairly reflect the relevant information of each layer by upper bounding layer-wise information. Therefore, CoIBA guarantees that the discarded activation is unnecessary in every targeted layer to make a decision. The extensive experimental results demonstrate the enhancement in faithfulness of the feature attributions provided by CoIBA.
****
OpticalNet: An Optical Imaging Dataset and Benchmark Beyond the Diffraction Limit
Benquan Wang, Ruyi An, Jin-Kyu So, Sergei Kurdiumov, Eng Aik Chan, Giorgio Adamo, Yuhan Peng, Yewen Li, Bo An; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10900-10912
Optical imaging capable of resolving nanoscale features would revolutionize scientific research and engineering applications across biomedicine, smart manufacturing, and semiconductor quality control. However, due to the physical phenomenon of diffraction, the optical resolution is limited to approximately half the wavelength of light, which impedes the observation of subwavelength objects such as the native state coronavirus, typically smaller than 200 nm. Fortunately, deep learning methods have shown remarkable potential in uncovering underlying patterns within data, promising to overcome the diffraction limit by revealing the mapping pattern between diffraction images and their corresponding ground truth object images. However, the absence of suitable datasets has hindered progress in this field--collecting high-quality optical data of subwavelength objects is highly difficult as these objects are inherently invisible under conventional microscopy, making it impossible to perform standard visual calibration and drift correction. Therefore, we provide the first general optical imaging dataset based on the "building block" concept for challenging the diffraction limit. Drawing an analogy to modular construction principles, we construct a comprehensive optical imaging dataset comprising subwavelength fundamental elements, i.e., small square units that can be assembled into larger and more complex objects. We then frame the task as an image-to-image translation task and evaluate various vision methods. Experimental results validate our "building block" concept, demonstrating that models trained on basic square units can effectively generalize to realistic, more complex unseen objects. Most importantly, by highlighting this underexplored AI-for-science area and its potential, we aspire to advance optical science by fostering collaboration with the vision and machine learning communities.
****
Dataset Distillation with Neural Characteristic Function: A Minmax Perspective
Shaobo Wang, Yicun Yang, Zhiyuan Liu, Chenghao Sun, Xuming Hu, Conghui He, Linfeng Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25570-25580
Dataset distillation has emerged as a powerful approach for reducing data requir

ements in deep learning. Among various methods, distribution matching-based appr oaches stand out for their balance of computational efficiency and strong perfor mance. However, existing distance metrics used in distribution matching often fa il to accurately capture distributional differences, leading to unreliable measu res of discrepancy. In this paper, we reformulate dataset distillation as a minm ax optimization problem and introduce Neural Characteristic Function Discrepancy (NCFD), a comprehensive and theoretically grounded metric for measuring distrib utional differences. NCFD leverages the Characteristic Function (CF) to encapsul ate full distributional information, employing a neural network to optimize the sampling strategy for the CF's frequency arguments, thereby maximizing the discr epancy to enhance distance estimation. Simultaneously, we minimize the differenc e between real and synthetic data under this optimized NCFD measure. Our approac h, termed Neural Characteristic Function Matching (NCFM), inherently aligns the phase and amplitude of neural features in the complex plane for both real and sy nthetic data, achieving a balance between realism and diversity in synthetic sam ples. Experiments demonstrate that our method achieves significant performance g ains over state-of-the-art methods on both low- and high-resolution datasets. No tably, we achieve a 22.9% accuracy boost on ImageSquawk. Our method also reduces GPU memory usage by over 300x and achieves 20x faster processing speeds compare d to state-of-the-art methods. To the best of our knowledge, this is the first w ork to achieve lossless compression of CIFAR-100 on a single NVIDIA 2080 Ti GPU using only 2.3 GB of memory. The code for this work is publicly available at: ht tps://github.com/gszfwsb/NCFM.

****

Free-viewpoint Human Animation with Pose-correlated Reference Selection

Fa-Ting Hong, Zhan Xu, Haiyang Liu, Qinjie Lin, Luchuan Song, Zhixin Shu, Yang Z hou, Duygu Ceylan, Dan Xu; Proceedings of the Computer Vision and Pattern Recogn ition Conference (CVPR), 2025, pp. 26253-26262

Diffusion-based human animation aims to animate a human character based on a sou rce human image as well as driving signals such as a sequence of poses. Leveragi ng the generative capacity of diffusion model, existing approaches are able to g enerate high-fidelity poses, but struggle with significant viewpoint changes, es pecially in zoom-in/zoom-out scenarios where camera-character distance varies. T his limits the applications such as cinematic shot type plan or camera control. We propose a pose-correlated reference selection diffusion network, supporting s ubstantial viewpoint variations in human animation. Our key idea is to enable th e network to utilize multiple reference images as input, since significant viewp oint changes often lead to missing appearance details on the human body. To elim inate the computational cost, we first introduce a novel pose correlation module to compute similarities between non-aligned target and source poses, and then p ropose an adaptive reference selection strategy, utilizing the attention map to identify key regions for animation generation. To train our model, we curated a large dataset from public TED talks featuring varied shots of the same character , helping the model learn synthesis for different perspectives. Our experimental results show that with the same number of reference images, our model performs favorably compared to the current SOTA methods under large viewpoint change. We further show that the adaptive reference selection is able to choose the most re levant reference regions to generate humans under free viewpoints.

****

CORE4D: A 4D Human-Object-Human Interaction Dataset for Collaborative Object REa rrangement

Yun Liu, Chengwen Zhang, Ruofan Xing, Bingda Tang, Bowen Yang, Li Yi; Proceeding s of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17 69-1782

Understanding how humans cooperatively rearrange household objects is critical f or VR/AR and human-robot interaction. However, in-depth studies on modeling thes e behaviors are under-researched due to the lack of relevant datasets. We fill t his gap by presenting CORE4D, a novel large-scale 4D human-object-human interact ion dataset focusing on collaborative object rearrangement, which encompasses di verse compositions of various object geometries, collaboration modes, and 3D sce

nes. With 1K human-object-human motion sequences captured in the real world, we enrich CORE4D by contributing an iterative collaboration retargeting strategy to augment motions to a variety of novel objects. Leveraging this approach, CORE4D comprises a total of 11K collaboration sequences spanning 3K real and virtual object shapes. Benefiting from extensive motion patterns provided by CORE4D, we benchmark two tasks aiming at generating human-object interaction: human-object motion forecasting and interaction synthesis. Extensive experiments demonstrate the effectiveness of our collaboration retargeting strategy and indicate that CORE4D has posed new challenges to existing human-object interaction generation methodologies.

****

PillarHist: A Quantization-aware Pillar Feature Encoder based on Height-aware Histogram

Sifan Zhou, Zhihang Yuan, Dawei Yang, Xing Hu, Jian Qian, Ziyu Zhao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27336-27345

Real-time and high-performance 3D object detection plays a critical role in autonomous driving and robotics. Recent pillar-based 3D object detectors have gained significant attention due to their compact representation and low computational overhead, making them suitable for onboard deployment and quantization. However, existing pillar-based detectors still suffer from information loss along height dimension and large numerical distribution difference during pillar feature encoding (PFE), which severely limits their performance and quantization potential. To address above issue, we first unveil the importance of different input information during PFE and identify the height dimension as a key factor in enhancing 3D detection performance. Motivated by this observation, we propose a height-aware pillar feature encoder, called PillarHist. Specifically, PillarHist statistics the discrete distribution of points at different heights within one pillar. This simple yet effective design greatly preserves the information along the height dimension while significantly reducing the computation overhead of the PFE. Meanwhile, PillarHist also constrains the arithmetic distribution of PFE input to a stable range, making it quantization-friendly. Notably, PillarHist operates exclusively within the PFE stage to enhance performance, enabling seamless integration into existing pillar-based methods without introducing complex operations. Extensive experiments show the effectiveness of PillarHist in terms of both efficiency and performance.

****

POp-GS: Next Best View in 3D-Gaussian Splatting with P-Optimality

Joey Wilson, Marcelino Almeida, Sachit Mahajan, Martin Labrie, Maani Ghaffari, Omid Ghasemalizadeh, Min Sun, Cheng-Hao Kuo, Arnab Sen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3646-3655

In this paper, we present a novel algorithm for quantifying uncertainty and information gained within 3D Gaussian Splatting (3D-GS) through P-Optimality. While 3D-GS has proven to be a useful world model with high-quality rasterizations, it does not natively quantify uncertainty or information, posing a challenge for real-world applications such as 3D-GS SLAM. We propose to quantify information gain in 3D-GS by reformulating the problem through the lens of optimal experimental design, which is a classical solution widely used in literature. By restructuring information quantification of 3D-GS through optimal experimental design, we arrive at multiple solutions, of which T-Optimality and D-Optimality perform the best quantitatively and qualitatively as measured on two popular datasets. Additionally, we propose a block diagonal covariance approximation which provides a measure of correlation at the expense of a greater computation cost.

****

Empowering Vector Graphics with Consistently Arbitrary Viewing and View-dependent Visibility

Yidi Li, Jun Xiao, Zhengda Lu, Yiqun Wang, Haiyong Jiang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18531-18540

This work presents a novel text-to-vector graphics generation approach, Dream3DVG, allowing for arbitrary viewpoint viewing, progressive detail optimization, an

d view-dependent occlusion awareness. Our approach is a dual-branch optimizatio n framework, consisting of an auxiliary 3D Gaussian Splatting optimization branc h and a 3D vector graphics optimization branch. The introduced 3DGS branch can bridge the domain gaps between text prompts and vector graphics with more consis tent guidance. Moreover, 3DGS allows for progressive detail control by schedulin g classifier-free guidance, facilitating guiding vector graphics with coarse sha pes at the initial stages and finer details at later stages. We also improve th e view-dependent occlusions by devising a visibility-awareness rendering module. Extensive results on 3D sketches and 3D iconographies, demonstrate the superio rity of the method on different abstraction levels of details, cross-view consis tency, and occlusion-aware stroke culling.

****

Semantic and Expressive Variations in Image Captions Across Languages

Andre Ye, Sebastin Santy, Jena D. Hwang, Amy X. Zhang, Ranjay Krishna; Proceedin gs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2 9667-29679

Most vision-language models today are primarily trained on English image-text pa irs, with non-English pairs often filtered out. Evidence from cross-cultural psy chology suggests that this approach will bias models against perceptual modes ex hibited by people who speak other (non-English) languages. We investigate semant ic and expressive variation in image captions across different languages; we ana lyze both human-annotated datasets and model-produced captions. By analyzing cap tions across seven languages (English, French, German, Russian, Chinese, Japanes e, Korean) in high-quality image captioning datasets (Crossmodal and Visual Geno me), we find that multilingual caption sets tend to provide richer visual descri ptions than monolingual (including English-only) ones; multilingual sets contain 46.0% more objects, 66.1% more relationships, and 66.8% more attributes. We obs erve the same results with multilingual captions produced by LLaVA and the Googl e Vertex API: for example, compared to monolingual captions, they cover 21.9% mo re objects,18.8% more relations, and 20.1% more attributes. These suggest that, across a large number of samples, different languages bias people and models to focus on different visual concepts. Finally, we show that models trained on imag e-text data in one language perform distinctly better on that language's test se t. Our work points towards the potential value of training vision models on mult ilingual data sources to widen the range/variation of descriptive information th ose models are exposed to.

****

ATP-LLaVA: Adaptive Token Pruning for Large Vision Language Models

Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang, Yansong Tang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24972-2 4982

Large Vision Language Models (LVLMs) have achieved significant success across mu lti-modal tasks. However, the computational cost of processing long visual token s can be prohibitively expensive on resource-limited devices. Previous methods h ave identified redundancy in visual tokens within the Large Language Model (LLM) decoder layers and have mitigated this by pruning tokens using a pre-defined or fixed ratio, thereby reducing computational overhead. Nonetheless, we observe t hat the impact of pruning ratio varies across different LLM layers and instances (image-prompt pairs). Therefore, it is essential to develop a layer-wise and in stance-wise vision token pruning strategy to balance computational cost and mode l performance effectively. We propose ATP-LLaVA, a novel approach that adaptivel y determines instance-specific token pruning ratios for each LLM layer. Specific ally, we introduce an Adaptive Token Pruning (ATP) module, which computes the im portance score and pruning threshold based on input instance adaptively. The ATP module can be seamlessly integrated between any two LLM layers with negligible computational overhead. Additionally, we develop a Spatial Augmented Pruning (SA P) strategy that prunes visual tokens with both token redundancy and spatial mod eling perspectives. Our approach reduces the average token count by 75% while ma intaining performance, with only a minimal 1.9% degradation across seven widely used benchmarks.

****
ADD: Attribution-Driven Data Augmentation Framework for Boosting Image Super-Resolution
Ze-Yu Mi, Yu-Bin Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23101-23110
Data augmentation (DA) stands out as a powerful technique to enhance the generalization capabilities of deep neural networks across diverse tasks. However, in low-level vision tasks, DA remains rudimentary (i.e., vanilla DA), facing a critical bottleneck due to information loss. In this paper, we introduce a novel Calibrated Attribution Maps (CAM) to generate saliency masks, followed by two saliency-based DA methods-- Attribution-Driven Data augmentation (ADD) and ADD+--designed to address this issue. CAM leverages integrated gradients and incorporates two key innovations: a global feature detector and calibrated integrated gradients. Based on CAM and the proposed methods, we have two new insights for low-level vision tasks: (1) increasing pixel diversity, as seen in vanilla DA, can improve performance, and (2) focusing on salient features while minimizing the impact of irrelevant pixels, as seen in saliency-based DA, more effectively enhances model performance. Additionally, we find and highlight the key guiding principle for designing saliency-based DA: a wider spectrum of degradation patterns. Extensive experiments demonstrate the compatibility and consistency of our method, as well as the significant performance improvement across various SR tasks and networks.
****
DIFFER: Disentangling Identity Features via Semantic Cues for Clothes-Changing Person Re-ID
Xin Liang, Yogesh S Rawat; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13980-13989
Clothes-changing person re-identification (CC-ReID) aims to recognize individuals under different clothing scenarios. Current CC-ReID approaches either concentrate on modeling body shape using additional modalities including silhouette, pose, and body mesh, potentially causing the model to overlook other critical biometric traits such as gender, age, and style, or they incorporate supervision through additional labels that the model tries to disregard or emphasize, such as clothing or personal attributes. However, these annotations are discrete in nature and do not capture comprehensive descriptions. In this work, we propose DIFFER: Disentangle Identity Features From Entangled Representations, a novel adversarial learning method that leverages textual descriptions to disentangle identity features. Recognizing that image features inherently mix inseparable information, DIFFER introduces NBDetach, a mechanism designed for feature disentanglement by leveraging the separable nature of text descriptions as supervision. It partitions the feature space into distinct subspaces and, through gradient reversal layers, effectively separates identity-related features from non-biometric features. We evaluate DIFFER on 4 different benchmark datasets (LTCC, PRCC, CelebreID-Light, and CCVID) to demonstrate its effectiveness and provide state-of-the-art performance across all the benchmarks. DIFFER consistently outperforms the baseline method, with improvements in top-1 accuracy of 3.6% on LTCC, 3.4% on PRCC, 2.5% on CelebReID-Light, and 1% on CCVID. Our code can be found at https://github.com/xliangp/DIFFER.git.
****
HyperPose: Hypernetwork-Infused Camera Pose Localization and an Extended Cambridge Landmarks Dataset
Ron Ferens, Yosi Keller; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11547-11557
In this work, we propose HyperPose, which utilizes hypernetworks in absolute camera pose regressors. The inherent appearance variations in natural scenes, attributable to environmental conditions, perspective, and lighting, induce a significant domain disparity between the training and test datasets. This disparity degrades the precision of contemporary localization networks. To mitigate this, we advocate for incorporating hypernetworks into single-scene and multiscene camera pose regression models. During inference, the hypernetwork dynamically computes

adaptive weights for the localization regression heads based on the particular input image, effectively narrowing the domain gap. Using indoor and outdoor data sets, we evaluate the HyperPose methodology across multiple established absolute pose regression architectures. In particular, we introduce and share the Extended Cambridge Landmarks (ECL), which is a novel localization dataset, based on the Cambridge Landmarks dataset, showing it in multiple seasons with significantly varying appearance conditions. Our empirical experiments demonstrate that HyperPose yields notable performance enhancements for both single- and multi-scene architectures. We have made our source code, pre-trained models, and ECL dataset openly available.

****

Critic-V: VLM Critics Help Catch VLM Errors in Multimodal Reasoning

Di Zhang, Jingdi Lei, Junxian Li, Xunzhi Wang, Yujie Liu, Zonglin Yang, Jiatong Li, Weida Wang, Suorong Yang, Jianbo Wu, Peng Ye, Wanli Ouyang, Dongzhan Zhou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9050-9061

Vision-language models (VLMs) have shown remarkable advancements in multimodal reasoning tasks. However, they still often generate inaccurate or irrelevant responses due to issues like hallucinated image understandings or unrefined reasoning paths. To address these challenges, we introduce Critic-V, a novel framework inspired by the Actor-Critic paradigm to boost the reasoning capability of VLMs. This framework decouples the reasoning process and critic process by integrating two independent components: the Reasoner, which generates reasoning paths based on visual and textual inputs, and the Critic, which provides constructive critique to refine these paths. In this approach, the Reasoner generates reasoning responses according to text prompts, which can evolve iteratively as a policy based on feedback from the Critic. This interaction process was theoretically driven by a reinforcement learning framework where the Critic offers natural language critiques instead of scalar rewards, enabling more nuanced feedback to boost the Reasoner's capability on complex reasoning tasks. The Critic model is trained using Direct Preference Optimization (DPO), leveraging a preference dataset of critiques ranked by Rule-based Reward (RBR) to enhance its critic capabilities. Evaluation results show that the Critic-V framework significantly outperforms existing methods, including GPT-4V, on 5 out of 8 benchmarks, especially regarding reasoning accuracy and efficiency. Combining a dynamic text-based policy for the Reasoner and constructive feedback from the preference-optimized Critic enables a more reliable and context-sensitive multimodal reasoning process. Our approach provides a promising solution to enhance the reliability of VLMs, improving their performance in real-world reasoning-heavy multimodal applications such as autonomous driving and embodied intelligence.

****

Mono3DVLT: Monocular-Video-Based 3D Visual Language Tracking

Hongkai Wei, Yang Yang, Shijie Sun, Mingtao Feng, Xiangyu Song, Qi Lei, Hongli Hu, Rong Wang, Huansheng Song, Naveed Akhtar, Ajmal Saeed Mian; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13886-13896

Visual-Language Tracking (VLT) is emerging as a promising paradigm to bridge the human-machine performance gap. For single objects, VLT broadens the problem scope to text-driven video comprehension. Yet, this direction is still confined to 2D spatial extents, currently lacking the ability to deal with 3D tracking in the confines of monocular video. Unfortunately, advances in 3D tracking mainly rely on expensive sensor inputs, e.g., point clouds, depth measurements, radar. Absence of language counterpart for the outputs of these mildly democratized sensors in the literature also hinders VLT expansion to 3D tracking. Addressing that, we make the first attempt towards extending VLT to 3D tracking based on monocular video. We present a comprehensive framework, introducing (i) the Monocular-Video-based 3D Visual Language Tracking (Mono3DVLT) task, (ii) a large-scale dataset for the task, called Mono3DVLT-V2X, and (iii) a customized neural model for the task. Our dataset is carefully curated, leveraging a Large Langauge Model (LLM) followed by human verification, composing natural language descriptions for 7

9,158 video sequences aiming at single object tracking, providing 2D and 3D boun ding box annotations. Our neural model, termed Mono3DVLT-MT, is the first target ed approach for the Mono3DVLT task. Comprising the pipeline of multi-modal feat ure extractor, visual-language encoder, tracking decoder and a tracking head, ou r model sets a strong baseline for the task on Mono3DVLT-V2X. Experimental resul ts show that our method significantly outperforms existing techniques on the Mon o3DVLT-V2X dataset. Our dataset and code are available in https://github.com/hon gkai-wei/Mono3DVLT.
****

Towards Universal Dataset Distillation via Task-Driven Diffusion
Ding Qi, Jian Li, Junyao Gao, Shuguang Dou, Ying Tai, Jianlong Hu, Bo Zhao, Yabi ao Wang, Chengjie Wang, Cairong Zhao; Proceedings of the Computer Vision and Pat tern Recognition Conference (CVPR), 2025, pp. 10557-10566
Dataset distillation (DD) condenses key information from large-scale datasets in to smaller synthetic datasets, reducing storage and computational costs for trai ning networks. However, recent research has primarily focused on image classific ation tasks, with limited expansion to detection and segmentation. Two key chall enges remain: (i) Task Optimization Heterogeneity, where existing methods focus on class-level information and fail to address the diverse needs of detection an d segmentation and (ii) Inflexible Image Generation, where current generation me thods rely on global updates for single-class targets and lack localized optimiz ation for specific object regions.To address these challenges, we propose a univ ersal dataset distillation framework, named UniDD, a task-driven diffusion model for diverse DD tasks, as illustrated in Fig.1. Our approach operates in two sta ges: Universal Task Knowledge Mining, which captures task-relevant information t hrough task-specific proxy model training, and Universal Task-Driven Diffusion, where these proxies guide the diffusion process to generate task-specific synthe tic images.Extensive experiments across ImageNet-1K, Pascal VOC, and MS COCO dem onstrate that UniDD consistently outperforms state-of-the-art methods. In partic ular, on ImageNet-1K with IPC-10, UniDD surpasses previous diffusion-based metho ds by 6.1%, while also reducing deployment costs.
****

Parametric Point Cloud Completion for Polygonal Surface Reconstruction
Zhaiyu Chen, Yuqing Wang, Liangliang Nan, Xiao Xiang Zhu; Proceedings of the Com puter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11749-11758
Existing polygonal surface reconstruction methods heavily depend on input comple teness and struggle with incomplete point clouds. We argue that while current po int cloud completion techniques may recover missing points, they are not optimiz ed for polygonal surface reconstruction, where the parametric representation of underlying surfaces remains overlooked. To address this gap, we introduce parame tric completion, a novel paradigm for point cloud completion, which recovers par ametric primitives instead of individual points to convey high-level geometric s tructures. Our presented approach, PaCo, enables high-quality polygonal surface reconstruction by leveraging plane proxies that encapsulate both plane parameter s and inlier points, proving particularly effective in challenging scenarios wit h highly incomplete data. Comprehensive evaluations of our approach on the ABC d ataset establish its effectiveness with superior performance and set a new stand ard for polygonal surface reconstruction from incomplete data. Project page: htt ps://parametric-completion.github.io.
****

SyncSDE: A Probabilistic Framework for Diffusion Synchronization
Hyunjun Lee, Hyunsoo Lee, Sookwan Han; Proceedings of the Computer Vision and Pa ttern Recognition Conference (CVPR), 2025, pp. 17508-17517
There have been many attempts to leverage multiple diffusion models for collabor ative generation, extending beyond the original domain. A prominent approach inv olves synchronizing multiple diffusion trajectories by mixing the estimated scor es to artificially correlate the generation processes. However, existing methods rely on naive heuristics, such as averaging, without considering task specifici ty. These approaches do not clarify why such methods work and often fail when a heuristic suitable for one task is blindly applied to others. In this paper, we

present a probabilistic framework for analyzing why diffusion synchronization wo rks and reveal where heuristics should be focused--modeling correlations between multiple trajectories and adapting them to each specific task. We further ident ify optimal correlation models per task, achieving better results than previous approaches that apply a single heuristic across all tasks without justification.
****

MaRI: Material Retrieval Integration across Domains
Jianhui Wang, Zhifei Yang, Yangfan He, Huixiong Zhang, Yuxuan Chen, Jingwei Huan g; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5814-5823
Accurate material retrieval is critical for creating realistic 3D assets. Existi ng methods rely on datasets that capture shape-invariant and lighting-varied rep resentations of materials, which are scarce and face challenges due to limited d iversity and inadequate real-world generalization. Most current approaches adopt traditional image search techniques. They fall short in capturing the unique pr operties of material spaces, leading to suboptimal performance in retrieval task s. Addressing these challenges, we introduce MaRI, a framework designed to bridg e the feature space gap between synthetic and real-world materials. MaRI constru cts a shared embedding space that harmonizes visual and material attributes thro ugh a contrastive learning strategy by jointly training a image and a material e ncoder, bringing similar materials and images closer while separating dissimilar pairs within the feature space. To support this, we construct a comprehensive d ataset comprising high-quality synthetic materials rendered with controlled shap e variations and diverse lighting conditions, along with real-world materials pr ocessed and standardized using material transfer techniques. Extensive experimen ts demonstrate the superior performance, accuracy, and generalization capabiliti es of MaRI across diverse and complex material retrieval tasks, outperforming ex isting methods.
****

MCCD: Multi-Agent Collaboration-based Compositional Diffusion for Complex Text-t o-Image Generation
Mingcheng Li, Xiaolu Hou, Ziyang Liu, Dingkang Yang, Ziyun Qian, Jiawei Chen, Ji njie Wei, Yue Jiang, Qingyao Xu, Lihua Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13263-13272
Diffusion models have shown excellent performance in text-to-image generation. H owever, existing methods often suffer from performance bottlenecks when dealing with complex prompts involving multiple objects, characteristics, and relations. Therefore, we propose a Multi-agent Collaboration-based Compositional Diffusio n (MCCD) for text-to-image generation for complex scenes. Specifically, we desig n a multi-agent collaboration based scene parsing module that generates an agent system containing multiple agents with different tasks using MLLMs to adequatel y extract multiple scene elements. In addition, Hierarchical Compositional diffu sion utilizes Gaussian mask and filtering to achieve the refinement of bounding box regions and highlights objects through region enhancement for accurate and h igh-fidelity generation of complex scenes. Comprehensive experiments demonstrate that our MCCD significantly improves the performance of the baseline models in a training-free manner, which has a large advantage in complex scene generation. The code will be open-source on github.
****

Dual Semantic Guidance for Open Vocabulary Semantic Segmentation
Zhengyang Wang, Tingliang Feng, Fan Lyu, Fanhua Shang, Wei Feng, Liang Wan; Proc eedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20212-20222
Open-vocabulary semantic segmentation aims to enable models to segment arbitrary categories. Currently, though pre-trained Vision-Language Models (VLMs) like CL IP have established a robust foundation for this task by learning to match text and image representations from large-scale data, their lack of pixel-level recog nition necessitates further fine-tuning. Most existing methods leverage text as a guide to achieve pixel-level recognition. However, the inherent biases in text semantic descriptions and the lack of pixel-level supervisory information make

it challenging to fine-tune CLIP-based models effectively. This paper considers leveraging image-text data to simultaneously capture the semantic information contained in both image and text, thereby constructing Dual Semantic Guidance and corresponding pixel-level pseudo annotations. Particularly, the visual semantic guidance is enhanced via explicitly exploring foreground regions and minimizing the influence of background. The dual semantic guidance is then jointly utilized to fine-tune CLIP-based segmentation models, achieving decent fine-grained recognition capabilities. As the comprehensive evaluation shows, our method outperforms state-of-art results with large margins, on eight commonly used datasets with/without background.

****

CroCoDL: Cross-device Collaborative Dataset for Localization

Hermann Blum, Alessandro Mercurio, Joshua O'Reilly, Tim Engelbracht, Mihai Dusmanu, Marc Pollefeys, Zuria Bauer; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27424-27434

Accurate localization plays a pivotal role in the autonomy of systems operating in unfamiliar environments, particularly when interaction with humans is expected. High-accuracy visual localization systems encompass various components, such as image retrievers, feature extractors, matchers, reconstruction and pose estimation methods. This complexity translates to the necessity of robust evaluation settings and pipelines. However, existing datasets and benchmarks primarily focus on single-agent scenarios, overlooking the critical issue of cross-device localization. Different agents with different sensors will show their own specific strengths and weaknesses, and the data they have available varies substantially. This work addresses this gap by enhancing an existing augmented reality visual localization benchmark with data from legged robots, and evaluating human-robot, cross-device mapping and localization. Our contributions extend beyond device diversity and include high environment variability, spanning ten distinct locations ranging from disaster sites to art exhibitions. Each scene in our dataset features recordings from robot agents, hand-held and head-mounted devices, and high-accuracy ground truth LiDAR scanners, resulting in a comprehensive multi-agent dataset and benchmark. This work represents a significant advancement in the field of visual localization benchmarking, with key insights into the performance of cross-device localization methods across diverse settings.

****

Q-Bench-Video: Benchmark the Video Quality Understanding of LMMs

Zicheng Zhang, Ziheng Jia, Haoning Wu, Chunyi Li, Zijian Chen, Yingjie Zhou, Wei Sun, Xiaohong Liu, Xiongkuo Min, Weisi Lin, Guangtao Zhai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3229-3239

With the rising interest in research on Large Multi-modal Models (LMMs) for video understanding, many studies have emphasized general video comprehension capabilities, neglecting the systematic exploration into video quality understanding. To address this oversight, we introduce Q-Bench-Video in this paper, a new benchmark specifically designed to evaluate LMMs' proficiency in discerning video quality. a) To ensure video source diversity, Q-Bench-Video encompasses videos from natural scenes, AI-generated Content (AIGC), and Computer Graphics (CG). b) Building on the traditional multiple-choice questions format with the Yes-or-No and What-How categories, we include Open-ended questions to better evaluate complex scenarios. Additionally, we incorporate the video pair quality comparison question to enhance comprehensiveness. c) Beyond the traditional Technical, Aesthetic, and Temporal distortions, we have expanded our evaluation aspects to include the dimension of AIGC distortions, which addresses the increasing demand for video generation. Finally, we collect a total of 2,378 question-answer pairs and test them on 12 open-source & 5 proprietary LMMs. Our findings indicate that while LMMs have a foundational understanding of video quality, their performance remains incomplete and imprecise, with a notable discrepancy compared to human performance. Through Q-Bench-Video, we seek to catalyze community interest, stimulate further research, and unlock the untapped potential of LMMs to close the gap in video quality understanding.

****

Glossy Object Reconstruction with Cost-effective Polarized Acquisition
Bojian Wu, Yifan Peng, Ruizhen Hu, Xiaowei Zhou; Proceedings of the Computer Vis ion and Pattern Recognition Conference (CVPR), 2025, pp. 422-431
The challenge of image-based 3D reconstruction for glossy objects lies in separa ting diffuse and specular components on glossy surfaces from captured images, a task complicated by the ambiguity in discerning lighting conditions and material properties using RGB data alone. While state-of-the-art methods rely on tailore d and/or high-end equipment for data acquisition, which can be cumbersome and t ime-consuming, this work introduces a scalable polarization-aided approach that employs cost-effective acquisition tools. By attaching a linear polarizer to rea dily available RGB cameras, multi-view polarization images can be captured witho ut the need for advance calibration or precise measurements of the polarizer ang le, substantially reducing system construction costs. The proposed approach repr esents polarimetric BRDF, Stokes vectors, and polarization states of object surf aces as neural implicit fields. These fields, combined with the polarizer angle, are retrieved by optimizing the rendering loss of input polarized images. By le veraging fundamental physical principles for the implicit representation of pola rization rendering, our method demonstrates superiority over existing techniques through experiments in public datasets and real captured images on both reconst ruction and novel view synthesis.
****

Generalizable Object Keypoint Localization from Generative Priors
Dongkai Wang, Jiang Duan, Liangjian Wen, Shiyu Xuan, Hao Chen, Shiliang Zhang; P roceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 202 5, pp. 20265-20274
Generalizable object keypoint localization is a fundamental computer vision task in understanding the object structure. It is challenging for existing keypoint localization methods because their limited training data cannot provide generali zable shape and semantic cues, leading to inferior performance and generalizatio n capability. Instead of relying on large scale training data, this work tackles this challenge by exploiting the rich priors from large generative models. We p ropose a data-efficient generalizable localization method named GenLoc. GenLoc e xtracts the generative priors from a pre-trained image generation model by calcu lating the correlation map between image latent feature and condition embedding. Those priors are hence optimized with our proposed heatmap expectation loss to perform object keypoint localization. Benefited by the rich knowledge of generat ive priors in understanding of object semantics and structures, GenLoc achieves superior performance on various object keypoint localization benchmarks. It show s more substantial performance enhancements in cross-domain, few-shot and zero-s hot evaluation settings, e.g., getting 20%+ AP enhancement over CLAMP in various zero-shot settings.
****

CLIP is Almost All You Need: Towards Parameter-Efficient Scene Text Retrieval wi thout OCR
Xugong Qin, Peng Zhang, Jun Jie Ou Yang, Gangyan Zeng, Yubo Li, Yuanyuan Wang, W anqian Zhang, Pengwen Dai; Proceedings of the Computer Vision and Pattern Recogn ition Conference (CVPR), 2025, pp. 24873-24883
Scene Text Retrieval (STR) seeks to identify all images containing a given query string. Existing methods typically rely on an explicit Optical Character Recogn ition (OCR) process of text spotting or localization, which is susceptible to co mplex pipelines and accumulated errors. To settle this, we resort to the Contras tive Language-Image Pre-training (CLIP) models, which have demonstrated the capa city to perceive and understand scene text, making it possible to achieve strict ly OCR-free STR. From the perspective of parameter-efficient transfer learning, a lightweight visual position adapter is proposed to provide a positional inform ation complement for the visual encoder. Besides, we introduce a visual context dropout technique to improve the alignment of local visual features. A novel, pa rameter-free cross-attention mechanism transfers the contrastive relationship be tween images and text to that between visual tokens and text, producing a rich c ross-modal representation, which can be utilized for efficient reranking with a

linear classifier. The resulting model, CAYN, which proves that CLIP is Almost a ll You Need for STR with no more than 0.50M additional parameters required, achi eves new state-of-the-art performance on the STR task, with 92.46%/89.49%/85.98% mAP on the SVT/IIIT-STR/TTR datasets. Our findings demonstrate that CLIP can se rve as a reliable and efficient solution for OCR-free STR.
****

L-SWAG: Layer-Sample Wise Activation with Gradients Information for Zero-Shot NA S on Vision Transformers
Sofia Casarin, Sergio Escalera, Oswald Lanz; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4441-4451
Training-free Neural Architecture Search (NAS) efficiently identifies high-perfo rming neural networks using zero-cost (ZC) proxies. Unlike multi-shot and one-sh ot NAS approaches, ZC-NAS is both (i) time-efficient, eliminating the need for m odel training, and (ii) interpretable, with proxy designs often theoretically gr ounded. Despite rapid developments in the field, current SOTA ZC proxies are typ ically constrained to well-established convolutional search spaces. With the ris e of Large Language Models shaping the future of deep learning, this work extend s ZC proxy applicability to Vision Transformers (ViTs). We present a new benchma rk using the Autoformer search space evaluated on 6 distinct tasks, and propose Layer-Sample Wise Activation with Gradients information (L-SWAG), a novel, gener alizable metric that characterises both convolutional and transformer architectu res across 14 tasks. Additionally, previous works highlighted how different prox ies contain complementary information, motivating the need for a ML model to ide ntify useful combinations. To further enhance ZC-NAS, we therefore introduce LIB RA-NAS (Low Information gain and Bias Re-Alignment), a method that strategically combines proxies to best represent a specific benchmark. Integrated into the NA S search, LIBRA-NAS outperforms evolution and gradient-based NAS techniques by i dentifying an architecture with a 17.0% test error on ImageNet1k in just 0.1 GPU days.
****

Commonsense Video Question Answering through Video-Grounded Entailment Tree Reas oning
Huabin Liu, Filip Ilievski, Cees G. M. Snoek; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3262-3271
This paper proposes the first video-grounded entailment tree reasoning method fo r commonsense video question answering (VQA). Despite the remarkable progress of large visual-language models (VLMs), there are growing concerns that they learn spurious correlations between videos and likely answers, reinforced by their bl ack-box nature and remaining benchmarking biases. Our method explicitly grounds VQA tasks to video fragments in four steps: entailment tree construction, video-language entailment verification, tree reasoning, and dynamic tree expansion. A vital benefit of the method is its generalizability to current video- and image-based VLMs across reasoning types.To support fair evaluation, we devise a de-bia sing procedure based on large-language models that rewrite VQA benchmark answer sets to enforce model reasoning. Systematic experiments on existing and de-biase d benchmarks highlight the impact of our method components across benchmarks, VL Ms, and reasoning types.
****

What Makes a Good Dataset for Knowledge Distillation?
Logan Frank, Jim Davis; Proceedings of the Computer Vision and Pattern Recogniti on Conference (CVPR), 2025, pp. 23755-23764
Knowledge distillation (KD) has been a popular and effective method for model co mpression. One important assumption of KD is that the teacher's original dataset will also be available when training the student. However, in situations such a s continual learning and distilling large models trained on company-withheld dat asets, having access to the original data may not always be possible. This leads practitioners towards utilizing other sources of supplemental data, which could yield mixed results. One must then ask: "what makes a good dataset for transfer ring knowledge from teacher to student?" Many would assume that only real in-dom ain imagery is viable, but is that the only option? In this work, we explore mul

tiple possible surrogate distillation datasets and demonstrate that many differe nt datasets, even unnatural synthetic imagery, can serve as a suitable alternati ve in KD. From examining these alternative datasets, we identify and present var ious criteria describing what makes a good dataset for distillation. Source code is available at https://github.com/osu-cvl/good-kd-dataset.

****

Lifelong Knowledge Editing for Vision Language Models with Low-Rank Mixture-of-E xperts

Qizhou Chen, Chengyu Wang, Dakan Wang, Taolin Zhang, Wangyue Li, Xiaofeng He; Pr oceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025 , pp. 9455-9466

Model editing aims to correct inaccurate knowledge, update outdated information, and incorporate new data into Large Language Models (LLMs) without the need for retraining. This task poses challenges in lifelong scenarios where edits must b e continuously applied for real-world applications. While some editors demonstra te strong robustness for lifelong editing in pure LLMs, Vision LLMs (VLLMs), whi ch incorporate an additional vision modality, are not directly adaptable to exis ting LLM editors. In this paper, we propose LiveEdit, a lifelong vision language model edit to bridge the gap between lifelong LLM editing and VLLMs. We begin b y training an editing expert generator to independently produce low-rank experts for each editing instance, with the goal of correcting the relevant responses o f the VLLM. A hard filtering mechanism is developed to utilize visual semantic k nowledge, thereby coarsely eliminating visually irrelevant experts for input que ries during the inference stage of the post-edited model. Finally, to integrate visually relevant experts, we introduce a soft routing mechanism based on textua l semantic relevance to achieve multi-expert fusion. For evaluation, we establis h a benchmark for lifelong VLLM editing. Extensive experiments demonstrate that LiveEdit offers significant advantages in lifelong VLLM editing scenarios. Furth er experiments validate the rationality and effectiveness of each module design in LiveEdit.

****

Rectification-specific Supervision and Constrained Estimator for Online Stereo R ectification

Rui Gong, Kim-Hui Yap, Weide Liu, Xulei Yang, Jun Cheng; Proceedings of the Comp uter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22348-22358

Online stereo rectification is critical for autonomous vehicles and robots in dy namic environments, where factors such as vibration, temperature fluctuations, a nd mechanical stress can affect rectification accuracy and severely degrade down stream stereo depth estimation. Current dominant approaches for online stereo re ctification involve estimating relative camera poses in real time to derive rect ification homographies. However, they do not directly optimize for rectification constraints. Additionally, the general-purpose correspondence matchers used in these methods are not trained for rectification, while training of these matcher s typically requires ground-truth correspondences which are not available in ste reo rectification datasets. To address these limitations, we propose a matching-based stereo rectification framework that is directly optimized for rectificatio n and does not require ground-truth correspondence annotations for training. We assume intrinsics are known as they are generally available on modern devices an d are relatively stable. Our framework incorporates a rectification-constrained estimator and applies multi-level, rectification-specific supervision that train s the matcher network for rectification without relying on ground-truth correspo ndences. Additionally, we create a new rectification dataset with ground-truth o ptical flow annotations, eliminating bias from evaluation metrics used in prior work that relied on pretrained keypoint matching or optical flow models. Extensi ve experiments show that our approach outperforms both state-of-the-art matching -based and matching-free methods in vertical flow metric by 10.7% on the Carla-F lowguided dataset and 21.3% on the Semi-Truck Highway dataset, offering superior rectification accuracy.

****

Shape and Texture: What Influences Reliable Optical Flow Estimation?

Libo Long, Xiao Hu, Jochen Lang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27894-27903

Recent methods have made significant progress in optical flow estimation. However, the evaluation of these methods focuses mainly on improved accuracy in benchmarks and often overlooks the analysis of network robustness, which may be important in safety-critical scenarios such as autonomous driving. In this paper, we propose a novel method for robustness evaluation by modifying data from original benchmarks. Unlike previous benchmarks that focus on complex scenes, we propose to modify shape and texture of objects from the original images in order to analyze the sensitivity to these changes observed in the output. Our aim is to identify common failure cases of state-of-the-art (SOTA) methods to evaluate their robustness and understand their behaviors. We show that: Optical flow methods are more sensitive to shape changes than to texture changes; and optical flow methods tend to "remember" objects seen during training and may "ignore" the motion of unseen objects. Our experimental results and findings provide a more in-depth understanding of the behavior of recent optical flow methods.

****

PartGen: Part-level 3D Generation and Reconstruction with Multi-view Diffusion Models

Minghao Chen, Roman Shapovalov, Iro Laina, Tom Monnier, Jianyuan Wang, David Novotny, Andrea Vedaldi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5881-5892

Text- or image-to-3D generators and 3D scanners can now produce 3D assets with high-quality shapes and textures, but as single, fused entities lacking meaningful structure. In contrast, most applications and creative workflows require 3D assets to be composed of distinct, meaningful parts that can be independently manipulated. To bridge this gap, we introduce PartGen, a novel approach for generating, from text, images, or unstructured 3D objects, 3D objects composed of meaningful parts. Our method leverages a multi-view diffusion model to extract plausible and view-consistent part segmentations from multiple views of a 3D object, dividing it into meaningful components. A second multi-view diffusion model then processes each part individually, filling in occlusions and generating completed views, which are subsequently passed to a 3D reconstruction network. The completion process ensures that the reconstructed parts integrate cohesively by considering the context of the entire object, compensating for missing information caused by occlusions and, in extreme cases, hallucinating entirely invisible parts based on contextual cues. We evaluate PartGen on both generated and real 3D assets, demonstrating significant improvements over segmentation and part completion baselines. We also showcase downstream applications such as text-guided 3D part editing.

****

FedCALM: Conflict-aware Layer-wise Mitigation for Selective Aggregation in Deeper Personalized Federated Learning

Hao Zheng, Zhigang Hu, Liu Yang, Meiguang Zheng, Aikun Xu, Boyu Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15444-15453

Server aggregation conflict is a key challenge in personalized federated learning (PFL). While existing PFL methods have achieved significant progress with shallow base models (e.g., four-layer CNNs), they often overlook the negative impacts of deeper base models on personalization mechanisms. In this paper, we identify the phenomenon of deep model degradation in PFL, where as base model depth increases, the model becomes more sensitive to local client data distributions, thereby exacerbating server aggregation conflicts and ultimately reducing overall model performance. Moreover, we show that these conflicts manifest in insufficient global average updates and mutual constraints between clients. Motivated by our analysis, we proposed a two-stage conflict-aware layer-wise mitigation algorithm (FedCALM), which first constructs a conflict-free global update to alleviate negative conflicts, and then maximizes the benefits of all clients through a conflict-aware strategy. Notably, our method naturally leads to a selective mechanism that balances the tradeoff between clients involved in aggregation and the to

lerance for conflicts. Consequently, it can boost the positive contribution to t
he clients even with the greatest conflicts with the global update. Extensive ex
periments across multiple datasets and deeper base models demonstrate that FedCA
LM outperforms four state-of-the-art (SOTA) methods by up to 9.88% and seamlessl
y integrates into existing PFL methods with performance improvements of up to 9.
01%.
****

SINR: Sparsity Driven Compressed Implicit Neural Representations
Dhananjaya Jayasundara, Sudarshan Rajagopalan, Yasiru Ranasinghe, Trac D. Tran,
Vishal M. Patel; Proceedings of the Computer Vision and Pattern Recognition Conf
erence (CVPR), 2025, pp. 3061-3070
Implicit Neural Representations (INRs) are increasingly recognized as a versatil
e data modality for representing discretized signals, offering benefits such as
infinite query resolution and reduced storage requirements. Existing signal comp
ression approaches for INRs typically employ one of two strategies: 1. direct qu
antization with entropy coding of the trained INR; 2. deriving a latent code on
top of the INR through a learnable transformation. Thus, their performance is he
avily dependent on the quantization and entropy coding schemes employed. In this
 paper, we introduce SINR, an innovative compression algorithm that leverages th
e patterns in the vector spaces formed by weights of INRs. We compress these vec
tor spaces using a high-dimensional sparse code within a dictionary. Further ana
lysis reveals that the atoms of the dictionary used to generate the sparse code
do not need to be learned or transmitted to successfully recover the INR weights
. We demonstrate that the proposed approach can be integrated with any existing
INR-based signal compression technique. Our results indicate that SINR achieves
substantial reductions in storage requirements for INRs across various configura
tions, outperforming conventional INR-based compression baselines. Furthermore,
SINR maintains high-quality decoding across diverse data modalities, including i
mages, occupancy fields, and Neural Radiance Fields.
****

CaricatureBooth: Data-Free Interactive Caricature Generation in a Photo Booth
Zhiyu Qu, Yunqi Miao, Zhensong Zhang, Jifei Song, Jiankang Deng, Yi-Zhe Song; Pr
oceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025
, pp. 10815-10824
We present CaricatureBooth, a system that transforms caricature creation into a
simple interactive experience -- as easy as using a photo booth! A key challenge
 in caricature generation is two-fold: the scarcity of high-quality caricature d
ata and the difficulty in enabling precise creative control over the exaggeratio
n process while maintaining identity. Prior approaches either require large-scal
e caricature and photo data or lack intuitive mechanisms for users to guide the
deformation without losing identity. We address the data scarcity by synthesisin
g training data through Thin Plate Spline (TPS) deformation of standard face ima
ges. For creative control, we design a Bezier curve interface where users can ea
sily manipulate facial features, with these edits then driving TPS transformatio
ns at inference time. When combined with a pre-trained ID-preserving diffusion m
odel, our system maintains both identity preservation and creative flexibility.
Through extensive experiments, we demonstrate that CaricatureBooth achieves stat
e-of-the-art quality while making the joy of caricature creation as accessible a
s taking a photo -- just walk in and walk out with your personalised caricature!
 Code is available at https://github.com/WinKawaks/CaricatureBooth.
****

FlexGS: Train Once, Deploy Everywhere with Many-in-One Flexible 3D Gaussian Spla
tting
Hengyu Liu, Yuehao Wang, Chenxin Li, Ruisi Cai, Kevin Wang, Wuyang Li, Pavlo Mol
chanov, Peihao Wang, Zhangyang Wang; Proceedings of the Computer Vision and Patt
ern Recognition Conference (CVPR), 2025, pp. 16336-16345
3D Gaussian splatting (3DGS) has enabled various applications in 3D scene repres
entation and novel view synthesis due to its efficient rendering capabilities. H
owever, 3DGS demands significant GPU memory, limiting its use on devices with re
stricted computational resources. Previous approaches have focused on pruning le

ss important Gaussians, effectively compressing 3DGS but often requiring a fine-tuning stage and lacking adaptability for the specific memory needs of different devices. In this work, we present an elastic inference method for 3DGS. Given an input for the desired model size, our method selects and transforms a subset of Gaussians, achieving substantial rendering performance without additional fine-tuning. We introduce a tiny learnable module that controls Gaussian selection based on the input percentage, along with a transformation module that adjusts the selected Gaussians to complement the performance of the reduced model. Comprehensive experiments on ZipNeRF, MipNeRF and Tanks&Temples scenes demonstrate the effectiveness of our approach. Code will be publicly available.

****

Generalizing Deepfake Video Detection with Plug-and-Play: Video-Level Blending and Spatiotemporal Adapter Tuning

Zhiyuan Yan, Yandan Zhao, Shen Chen, Mingyi Guo, Xinghe Fu, Taiping Yao, Shouhong Ding, Yunsheng Wu, Li Yuan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12615-12625

Three key challenges hinder the development of current deepfake video detection: (1) Temporal features can be complex and diverse: how can we identify general temporal artifacts to enhance model generalization? (2) Spatiotemporal models often lean heavily on one type of artifact and ignore the other: how can we ensure balanced learning from both? (3) Videos are naturally resource-intensive: how can we tackle efficiency without compromising accuracy? This paper attempts to tackle the three challenges jointly. First, inspired by the notable generality of using image-level blending data for image forgery detection, we investigate whether and how video-level blending can be effective in video. We then perform a thorough analysis and identify a previously underexplored temporal forgery artifact: Facial Feature Drift (FFD), which commonly exists across different forgeries. To reproduce FFD, we then propose a novel Video-level Blending data (VB), where VB is implemented by blending the original image and its warped version frame-by-frame, serving as a hard negative sample to mine more general artifacts. Second, we carefully design a lightweight Spatiotemporal Adapter (StA) to equip a pre-trained image model with the ability to capture both spatial and temporal features jointly and efficiently. StA is designed with two-stream 3D-Conv with varying kernel sizes, allowing it to process spatial and temporal features separately. This eliminates the need to design a new deepfake-specific video architecture from scratch. Extensive experiments validate the effectiveness of the proposed methods; and show our approach can generalize well to previously unseen forgery videos.

****

ManipTrans: Efficient Dexterous Bimanual Manipulation Transfer via Residual Learning

Kailin Li, Puhao Li, Tengyu Liu, Yuyang Li, Siyuan Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6991-7003

Human hands play a central role in interacting, motivating increasing research in dexterous robotic manipulation. Data-driven embodied AI algorithms demand precise, large-scale, human-like manipulation sequences, which are challenging to obtain with conventional reinforcement learning or real-world teleoperation. To address this, we introduce ManipTrans, a novel two-stage method for efficiently transferring human bimanual skills to dexterous robotic hands in simulation. ManipTrans first pre-trains a generalist trajectory imitator to mimic hand motion, then fine-tunes a specific residual module under interaction constraints, enabling efficient learning and accurate execution of complex bimanual tasks. Experiments show that ManipTrans surpasses state-of-the-art methods in success rate, fidelity, and efficiency. Leveraging ManipTrans, we transfer multiple hand-object datasets to robotic hands, creating DexManipNet, a large-scale dataset featuring previously unexplored tasks like pen capping and bottle unscrewing. DexManipNet comprises 3.3K episodes of robotic manipulation and is easily extensible, facilitating further policy training for dexterous hands and enabling real-world deployments.

****

Precise, Fast, and Low-cost Concept Erasure in Value Space: Orthogonal Complement Matters
Yuan Wang, Ouxiang Li, Tingting Mu, Yanbin Hao, Kuien Liu, Xiang Wang, Xiangnan He; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28759-28768
The success of text-to-image generation enabled by diffusion models has imposed an urgent need to erase unwanted concepts, e.g., copyrighted, offensive, and unsafe ones, from the pre-trained models in a precise, timely, and low-cost manner. The twofold demand of concept erasure requires a precise removal of the target concept during generation (i.e., erasure efficacy), while a minimal impact on non-target content generation (i.e., prior preservation). Existing methods are either computationally costly or face challenges in maintaining an effective balance between erasure efficacy and prior preservation. To improve, we propose a precise, fast, and low-cost concept erasure method, called Adaptive Vaule Decomposer (AdaVD), which is training-free. This method is grounded in a classical linear algebraic orthogonal complement operation, implemented in the value space of each cross-attention layer within the UNet of diffusion models. An effective shift factor is designed to adaptively navigate the erasure strength, enhancing prior preservation without sacrificing erasure efficacy. Extensive experimental results show that the proposed AdaVD is effective at both single and multiple concept erasure, showing a 2- to 10-fold improvement in prior preservation as compared to the second best, meanwhile achieving the best or near best erasure efficacy, when comparing with both training-based and training-free state of the arts. AdaVD supports a series of diffusion models and downstream image generation tasks, with code to be publicly available.
****
HOIGen-1M: A Large-scale Dataset for Human-Object Interaction Video Generation
Kun Liu, Qi Liu, Xinchen Liu, Jie Li, Yongdong Zhang, Jiebo Luo, Xiaodong He, Wu Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24001-24010
Text-to-video (T2V) generation has made tremendous progress in generating complicated scenes based on texts. However, human-object interaction (HOI) often cannot be precisely generated by current T2V models due to the lack of large-scale videos with accurate captions for HOI. To address this issue, we introduce HOIGen-1M, the first large-scale dataset for HOI Generation, consisting of over one million high-quality videos collected from diverse sources. In particular, to guarantee the high quality of videos, we first design an efficient framework to automatically curate HOI videos using the powerful multimodal large language models (MLLMs), and then the videos are further cleaned by human annotators. Moreover, to obtain accurate textual captions for HOI videos, we design a novel video description method based on a Mixture-of-Multimodal-Experts (MoME) strategy that not only generates expressive captions but also eliminates the hallucination by individual MLLM. Furthermore, due to the lack of an evaluation framework for generated HOI videos, we propose two new metrics to assess the quality of generated videos in a coarse-to-fine manner. Extensive experiments reveal that current T2V models struggle to generate high-quality HOI videos and confirm that our HOIGen-1M dataset is instrumental for improving HOI video generation.
****
T2ISafety: Benchmark for Assessing Fairness, Toxicity, and Privacy in Image Generation
Lijun Li, Zhelun Shi, Xuhao Hu, Bowen Dong, Yiran Qin, Xihui Liu, Lu Sheng, Jing Shao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13381-13392
Text-to-image (T2I) models have rapidly advanced, enabling the generation of high-quality images from text prompts across various domains. However, these models present notable safety concerns, including the risk of generating harmful, biased, or private content. Current research on assessing T2I safety remains in its early stages. While some efforts have been made to evaluate models on specific safety dimensions, many critical risks remain unexplored. To address this gap, we introduce T2ISafety, a safety benchmark that evaluates T2I models across three

key domains: toxicity, fairness, and bias. We build a detailed hierarchy of 12 tasks and 44 categories based on these three domains, and meticulously collect 70K corresponding prompts. Based on this taxonomy and prompt set, we build a large-scale T2I dataset with 68K manually annotated images and train an evaluator capable of detecting critical risks that previous work has failed to identify, including risks that even ultra-large proprietary models like GPTs cannot correctly detect. We evaluate 15 prominent diffusion models on T2ISafety and reveal several concerns including persistent issues with racial fairness, a tendency to generate toxic content, and significant variation in privacy protection across the models, even with defense methods like concept erasing.

****

Order-One Rolling Shutter Cameras

Marvin Anas Hahn, Kathlén Kohn, Orlando Marigliano, Tomas Pajdla; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27007-27016

Rolling shutter (RS) cameras dominate consumer and smartphone markets. Several methods for computing the absolute pose of RS cameras have appeared in the last 20 years, but the relative pose problem has not been fully solved yet. We provide a unified theory for the important class of order-one rolling shutter (RS_1) cameras. These cameras generalize the perspective projection to RS cameras, projecting a generic space point to exactly one image point via a rational map. We introduce a new back-projection RS camera model, characterize RS_1 cameras, construct explicit parameterizations of such cameras, and determine the image of a space line. We classify all minimal problems for solving the relative camera pose problem with linear RS_1 cameras and discover new practical cases. Finally, we show how the theory can be used to explain RS models previously used for absolute pose computation.

****

Animate and Sound an Image

Xihua Wang, Ruihua Song, Chongxuan Li, Xin Cheng, Boyuan Li, Yihan Wu, Yuyue Wang, Hongteng Xu, Yunfeng Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23369-23378

This paper addresses a promising yet underexplored task, Image-to-Sounding-Video (I2SV) generation, which animates a static image and generates synchronized sound simultaneously. Despite advances in video and audio generation models, challenges remain to develop a unified model for generating naturally sounding videos. In this work, we propose a novel approach that leverages two separate pretrained diffusion models and makes vision and audio influence each other during generation based on the Diffusion Transformer (DiT) architecture. First, the individual video and audio pretrained generation models are decomposed into input, output, and expert sub-modules. We propose using a unified joint DiT block to integrate the expert sub-modules to effectively model the interaction between the two modalities, resulting in high-quality I2SV generation. Then, we introduce a joint classifier-free guidance technique to boost the performance during joint generation. Finally, we conduct extensive experiments on three popular benchmark datasets, and in both objective and subjective evaluation our method surpass all the baseline methods in almost all metrics. Case studies show our generated sounding videos are high quality and synchronized between video and audio.

****

Shining Yourself: High-Fidelity Ornaments Virtual Try-on with Diffusion Model

Yingmao Miao, Zhanpeng Huang, Rui Han, Zibin Wang, Chenhao Lin, Chao Shen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 359-368

While virtual try-on for clothes and shoes with diffusion models has gained attraction, virtual try-on for ornaments, such as bracelets, rings, earrings, and necklaces, remains largely unexplored. Due to the intricate tiny patterns and repeated geometric sub-structures in most ornaments, it is much more difficult to guarantee identity and appearance consistency under large pose and scale variances between ornaments and models. This paper proposes the task of virtual try-on for ornaments and presents a method to improve the geometric and appearance preser

vation of ornament virtual try-ons. Specifically, we estimate an accurate wearin g mask to improve the alignments between ornaments and models in an iterative sc heme alongside the denoising process. To preserve structure details, we further regularize attention layers to map the reference ornament mask to the wearing ma sk in an implicit way. Experiment results demonstrate that our method successful ly wears ornaments from reference images onto target models, handling substantia l differences in scale and pose while preserving identity and achieving realisti c visual effects.
****

Foveated Instance Segmentation
Hongyi Zeng, Wenxuan Liu, Tianhua Xia, Jinhui Chen, Ziyun Li, Sai Qian Zhang; Pr oceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025 , pp. 24496-24505
Instance segmentation is essential for augmented reality and virtual reality (AR /VR) as it enables precise object recognition and interaction, enhancing the int egration of virtual and real-world elements for an immersive experience. However , the high computational overhead of segmentation limits its application on reso urce-constrained AR/VR devices, causing large processing latency and degrading u ser experience. In contrast to conventional scenarios, AR/VR users typically fo cus on only a few regions within their field of view before shifting perspective , allowing segmentation to be concentrated on gaze-specific areas. This insight drives the need for efficient segmentation methods that prioritize processing in stance of interest, reducing computational load and enhancing real-time performa nce. In this paper, we present a foveated instance segmentation(FovealSeg) frame work that leverages real-time user gaze data to perform instance segmentation ex clusively on instance of interest, resulting in substantial computational saving s. Evaluation results show that FSNet achieves an IoU of 0.56 on ADE20K and 0.54 on LVIS, notably outperforming the baseline. The code is available at https://g ithub.com/SAI-Lab-NYU/Foveated-Instance-Segmentation.
****

Make It Count: Text-to-Image Generation with an Accurate Number of Objects
Lital Binyamin, Yoad Tewel, Hilit Segev, Eran Hirsch, Royi Rassin, Gal Chechik; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 20 25, pp. 13242-13251
Despite the unprecedented success of text-to-image diffusion models, controlling the number of depicted objects using text is surprisingly hard. This is importa nt for various applications from technical documents, to children's books to ill ustrating cooking recipes. Generating object-correct counts is fundamentally cha llenging because the generative model needs to keep a sense of separate identity for every instance of the object, even if several objects look identical or ove rlap, and then carry out a global computation implicitly during generation. It i s still unknown if such representations exist. To address count-correct generati on, we first identify features within the diffusion model that can carry the obj ect identity information. We then use them to separate and count instances of ob jects during the denoising process and detect over-generation and under-generati on. We fix the latter by training a model that predicts both the shape and locat ion of a missing object, based on the layout of existing ones, and show how it c an be used to guide denoising with correct object count. Our approach, CountGen, does not depend on external source to determine object layout, but rather uses the prior from the diffusion model itself, creating prompt-dependent and seed-de pendent layouts. Evaluated on two benchmark datasets, we find that CountGen stro ngly outperforms the count-accuracy of existing baselines.
****

Universal Domain Adaptation for Semantic Segmentation
Seun-An Choe, Keon-Hee Park, Jinwoo Choi, Gyeong-Moon Park; Proceedings of the C omputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4607-4617
Unsupervised domain adaptation for semantic segmentation (UDA-SS) aims to transf er knowledge from labeled synthetic data (source) to unlabeled real-world data (target). Traditional UDA-SS methods work on the assumption that the category set tings between the source and target domains are known in advance. However, in re

al-world scenarios, the category settings of the source and target are unknown d ue to the lack of labels in the target, resulting in the existence of target-pri vate or source-private classes. Traditional UDA-SS methods struggle with this ch ange, leading to negative transfer and performance degradation. To address these issues, we propose Universal Domain Adaptation for Semantic Segmentation (UniDA -SS) for the first time to achieve good performance even when the category setti ngs of source and target are unknown. We defined the problem in the UniDA-SS sce nario as that the confidence score of common Unsupervised domain adaptation for semantic segmentation (UDA-SS) aims to transfer knowledge from labeled source da ta to unlabeled target data. However, traditional UDA-SS methods assume that cat egory settings between source and target domains are known, which is unrealistic in real-world scenarios. This leads to performance degradation if private priva te classes exist. To address this limitation, we propose Universal Domain Adapta tion for Semantic Segmentation (UniDA-SS), achieving robust adaptation even with out prior knowledge of category settings. We define the problem in the UniDA-SS scenario as low confidence scores of common classes in the target domain, which leads to confusion with private classes. To solve this problem, we propose UniMA P: UniDA-SS with Image Ma tching and Prototype-based Distinction, a novel framew ork composed of two key components. First, Domain-Specific Prototype-based Disti nction (DSPD) divides each class into two domain-specific prototypes, enabling f iner separation of domain-specific features and enhancing the identification of common classes across domains. Second, Target-based Image Matching (TIM) selects a source image containing the most common-class pixels based on the target pseu do-label and pairs it in a batch to promote effective learning of common classes . We also introduce a new UniDA-SS benchmark and demonstrate through various exp eriments that UniMAP significantly outperforms baselines. The code is available at https://github.com/KU-VGI/UniDA-SS.

****

HyperGS: Hyperspectral 3D Gaussian Splatting

Christopher Thirgood, Oscar Mendez, Erin Ling, Jon Storey, Simon Hadfield; Proce edings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, p p. 5970-5979

We introduce HyperGS, a novel framework for Hyperspectral Novel View Synthesis (HNVS), based on a new latent 3D Gaussian Splatting (3DGS) technique. Our approac h enables simultaneous spatial and spectral renderings by encoding material prop erties from multi-view 3D hyperspectral datasets. HyperGS reconstructs high-fide lity views from arbitrary perspectives with improved accuracy and speed, outperf orming currently existing methods. To address the challenges of high-dimensional data, we perform view synthesis in a learned latent space, incorporating a pixe l-wise adaptive density function and a pruning technique for increased training stability and efficiency. Additionally, we introduce the first HNVS benchmark, i mplementing a number of new baselines based on recent SOTA RGB-NVS techniques, a longside the small number of prior works on HNVS. We demonstrate HyperGS's robus tness through extensive evaluation of real and simulated hyperspectral scenes wi th a 14dB accuracy improvement upon previously published models.

****

Emphasizing Discriminative Features for Dataset Distillation in Complex Scenario s

Kai Wang, Zekai Li, Zhi-Qi Cheng, Samir Khaki, Ahmad Sajedi, Ramakrishna Vedanta m, Konstantinos N Plataniotis, Alexander Hauptmann, Yang You; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30451-3046 1

Dataset distillation has demonstrated strong performance on simple datasets like CIFAR, MNIST, and TinyImageNet but struggles to achieve similar results in more complex scenarios. In this paper, we propose EDF (emphasizes the discriminative features), a dataset distillation method that enhances key discriminative regio ns in synthetic images using Grad-CAM activation maps. Our approach is inspired by a key observation: in simple datasets, high-activation areas typically occupy most of the image, whereas in complex scenarios, the size of these areas is muc h smaller. Unlike previous methods that treat all pixels equally when synthesizi

ng images, EDF uses Grad-CAM activation maps to enhance high-activation areas. From a supervision perspective, we downplay supervision signals produced by lower trajectory-matching losses, as they contain common patterns. Additionally, to help the DD community better explore complex scenarios, we build the Complex Dataset Distillation (Comp-DD) benchmark by meticulously selecting sixteen subsets, eight easy and eight hard, from ImageNet-1K. In particular, EDF consistently outperforms SOTA results in complex scenarios, such as ImageNet-1K subsets. Hopefully, more researchers will be inspired and encouraged to improve the practicality and efficacy of DD. Our code and benchmark have been made public at NUS-HPC-AI-Lab/EDF.

****

LMO: Linear Mamba Operator for MRI Reconstruction

Wei Li, Jiawei Jiang, Jie Wu, Kaihao Yu, Jianwei Zheng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5112-5122

Interpretability and consistency have long been crucial factors in MRI reconstruction. While interpretability has been significantly innovated with the emerging deep unfolding networks, current solutions still suffer from inconsistency issues and produce inferior anatomical structures. Especially in out-of-distribution cases, e.g., when the acceleration rate (AR) varies, the generalization performance is often catastrophic. To counteract the dilemma, we propose an innovative Linear Mamba Operator (LMO) to ensure consistency and generalization, while still enjoying desirable interpretability. Theoretically, we argue that mapping between function spaces, rather than between signal instances, provides a solid foundation of high generalization. Technically, LMO achieves a good balance between global integration facilitated by a state space model that scans the whole function domain, and local integration engaged with an appealing property of continuous-discrete equivalence. On that basis, learning holistic features can be guaranteed, tapping the potential of maximizing data consistency. Quantitative and qualitative results demonstrate that LMO significantly outperforms other state-of-the-arts. More importantly, LMO is the unique model that, with AR changed, achieves retraining performance without retraining steps. Codes are available at https://github.com/ZhengJianwei2/LMO.

****

AnomalyNCD: Towards Novel Anomaly Class Discovery in Industrial Scenarios

Ziming Huang, Xurui Li, Haotian Liu, Feng Xue, Yuzhe Wang, Yu Zhou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4755-4765

Recently, multi-class anomaly classification has garnered increasing attention. Previous methods directly cluster anomalies but often struggle due to the lack of anomaly-prior knowledge. Acquiring this knowledge faces two issues: the non-prominent and weak-semantics anomalies. In this paper, we propose AnomalyNCD, a multi-class anomaly classification network compatible with different anomaly detection methods. To address the non-prominence of anomalies, we design main element binarization (MEBin) to obtain anomaly-centered images, ensuring anomalies are learned while avoiding the impact of incorrect detections. Next, to learn anomalies with weak semantics, we design mask-guided representation learning, which focuses on isolated anomalies guided by masks and reduces confusion from erroneous inputs through corrected pseudo labels. Finally, to enable flexible classification at both region and image levels, we develop a region merging strategy that determines the overall image category based on the classified anomaly regions. Our method outperforms the state-of-the-art works on the MVTec AD and MTD datasets. Compared with the current methods, AnomalyNCD combined with zero-shot anomaly detection method achieves a 10.8% F1 gain, 8.8% NMI gain, and 9.5% ARI gain on MVTec AD, and 12.8% F1 gain, 5.7% NMI gain, and 10.8% ARI gain on MTD. Code is available at https://github.com/HUST-SLOW/AnomalyNCD.

****

Segment This Thing: Foveated Tokenization for Efficient Point-Prompted Segmentation

Tanner Schmidt, Richard Newcombe; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29428-29437

This paper presents Segment This Thing (STT), a new efficient image segmentation model designed to produce a single segment given a single point prompt. Instead of following prior work and increasing efficiency by decreasing model size, we gain efficiency by foveating input images. Given an image and a point prompt, we extract a crop centered on the prompt and apply a novel variable-resolution patch tokenization in which patches are downsampled at a rate that increases with increased distance from the prompt. This approach yields far fewer image tokens than uniform patch tokenization. As a result we can drastically reduce the computational cost of segmentation without reducing model size. Furthermore, the foveation focuses the model on the region of interest, a potentially useful inductive bias. We show that our Segment This Thing model is more efficient than prior work while remaining competitive on segmentation benchmarks. It can easily run at interactive frame rates on consumer hardware and is thus a promising tool for augmented reality or robotics applications.

****

Task-Specific Gradient Adaptation for Few-Shot One-Class Classification

Yunlong Li, Xiabi Liu, Liyuan Pan, Yuchen Ren; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30556-30565

Optimization-based meta-learning methods for few-shot one-class classification (FS-OCC) aim to fine-tune a meta-trained model to classify the positive and negative samples using only a few positive samples by adaptation. However, recent approaches primarily focus on adjusting existing meta-learning algorithms for FS-OCC, while overlooking issues stemming from the misalignment between the cross-entropy loss and OCC tasks during adaptation. This misalignment, combined with the limited availability of one-class samples and the restricted diversity of task-specific adaptation, can significantly exacerbate the adverse effects of gradient instability and generalization. To address these challenges, we propose a novel Task-Specific Gradient Adaptation (TSGA) for FS-OCC. Without extra supervision, TSGA learns to generate appropriate, stable gradients by leveraging label prediction and feature representation details of one-class samples and refines the adaptation process by recalibrating task-specific gradients and regularization terms. We evaluate TSGA on three challenging datasets and a real-world CNC Milling Machine application and demonstrate consistent improvements over baseline methods. Furthermore, we illustrate the critical impact of gradient instability and task-agnostic adaptation. Notably, TSGA achieves state-of-the-art results by effectively addressing these issues.

****

TraF-Align: Trajectory-aware Feature Alignment for Asynchronous Multi-agent Perception

Zhiying Song, Lei Yang, Fuxi Wen, Jun Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12048-12057

Cooperative perception presents significant potential for enhancing the sensing capabilities of individual vehicles, however, inter-agent latency remains a critical challenge. Latencies cause misalignments in both spatial and semantic features, complicating the fusion of real-time observations from the ego vehicle with delayed data from others. To address these issues, we propose TraF-Align, a novel framework that learns the flow path of features by predicting the feature-level trajectory of objects from past observations up to the ego vehicle's current time. By generating temporally ordered sampling points along these paths, TraF-Align directs attention from the current-time query to relevant historical features along each trajectory, supporting the reconstruction of current-time features and promoting semantic interaction across multiple frames. This approach corrects spatial misalignment and ensures semantic consistency across agents, effectively compensating for motion and achieving coherent feature fusion. Experiments on two real-world datasets, V2V4Real and DAIR-V2X-Seq, show that TraF-Align sets a new benchmark for asynchronous cooperative perception.

****

DreamCache: Finetuning-Free Lightweight Personalized Image Generation via Feature Caching

Emanuele Aiello, Umberto Michieli, Diego Valsesia, Mete Ozay, Enrico Magli; Proc

eedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12480-12489
Personalized image generation requires text-to-image generative models that capture the core features of a reference subject to allow for controlled generation across different contexts. Existing methods face challenges due to complex training requirements, high inference costs, limited flexibility, or a combination of these issues. In this paper, we introduce DreamCache, a scalable approach for efficient and high-quality personalized image generation. By caching a small number of reference image features from a subset of layers and a single timestep of the pretrained diffusion denoiser, DreamCache enables dynamic modulation of the generated image features through lightweight, trained conditioning adapters. DreamCache achieves state-of-the-art image and text alignment, utilizing an order of magnitude fewer extra parameters, and is both more computationally effective and versatile than existing models.
****

3D Gaussian Inpainting with Depth-Guided Cross-View Consistency
Sheng-Yu Huang, Zi-Ting Chou, Yu-Chiang Frank Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26704-26713
When performing 3D inpainting using novel-view rendering methods like Neural Radiance Field (NeRF) or 3D Gussian Splatting (3DGS), how to achieve texture and geometry consistency across camera views has been a challenge. In this paper, we propose a framework of 3D Gaussian Inpainting with Depth-Guided Cross-View Consistency (3DGIC) for cross-view consistent 3D inpainting. Guided by the rendered depth information from each training view, our 3DGIC exploits background pixels visible across different views for updating the inpainting mask, allowing us to refine the 3DGS for inpainting purposes. Through extensive experiments on benchmark datasets, we confirm that our 3DGIC outperforms current state-of-the-art 3D inpainting methods quantitatively and qualitatively.
****

Your Large Vision-Language Model Only Needs A Few Attention Heads For Visual Grounding
Seil Kang, Jinyeong Kim, Junhyeok Kim, Seong Jae Hwang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9339-9350
Visual grounding seeks to localize the image region corresponding to a free-form text description. Recently, the strong multimodal capabilities of Large Vision-Language Models (LVLMs) have driven substantial improvements in visual grounding, though they inevitably require fine-tuning and additional model components to explicitly generate bounding boxes or segmentation masks. However, we discover that a few attention heads in frozen LVLMs demonstrate strong visual grounding capabilities. We refer to these heads, which consistently capture object locations related to text semantics, as localization heads. Using localization heads, we introduce a straightforward and effective training-free visual grounding framework that utilizes text-to-image attention maps from localization heads to identify the target objects. Surprisingly, only three out of thousands of attention heads are sufficient to achieve competitive localization performance compared to existing LVLM-based visual grounding methods that require fine-tuning. Our findings suggest that LVLMs can innately ground objects based on a deep comprehension of the text-image relationship, as they implicitly focus on relevant image regions to generate informative text outputs.
****

FlexUOD: The Answer to Real-world Unsupervised Image Outlier Detection
Zhonghang Liu, Kun Zhou, Changshuo Wang, Wen-Yan Lin, Jiangbo Lu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15183-15193
How many outliers are within an unlabeled and contaminated dataset? Despite a series of unsupervised outlier detection (UOD) approaches have been proposed, they cannot correctly answer this critical question, resulting in their performance instability across various real-world (varying contamination factor) scenarios. To address this problem, we propose FlexUOD, with a novel contamination factor estimation perspective. FlexUOD not only achieves its remarkable robustness but a

lso is a general and plug-and-play framework, which can significantly improve th e performance of existing UOD methods. Extensive experiments demonstrate that Fl exUOD achieves state-of-the-art results as well as high efficacy on diverse eval uation benchmarks.
****

Ges3ViG : Incorporating Pointing Gestures into Language-Based 3D Visual Groundin g for Embodied Reference Understanding
Atharv Mahesh Mane, Dulanga Weerakoon, Vigneshwaran Subbaraju, Sougata Sen, Sanj ay E. Sarma, Archan Misra; Proceedings of the Computer Vision and Pattern Recogn ition Conference (CVPR), 2025, pp. 9017-9026
3-Dimensional Embodied Reference Understanding (3DERU) combines a language descr iption and an accompanying pointing gesture to identify the most relevant target object in a 3D scene. Although prior work has explored pure language-based 3D g rounding, there has been limited exploration of 3D-ERU, which also incorporates human pointing gestures. To address this gap, we introduce a data augmentation f ramework-Imputer, and use it to curate a new benchmark dataset-ImputeRefer for 3 D-ERU, by incorporating human pointing gestures into existing 3D scene datasets that only contain language instructions. We also propose Ges3ViG, a novel model for 3D-ERU that achieves 30% improvement in accuracy as compared to other 3DERU models and 9% compared to other purely language-based 3D grounding models. Our code and dataset are available at https://github.com/AtharvMane/Ges3ViG.
****

Focusing on Tracks for Online Multi-Object Tracking
Kyujin Shim, Kangwook Ko, Yujin Yang, Changick Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11687-11696
Multi-object tracking (MOT) is a critical task in computer vision, requiring the accurate identification and continuous tracking of multiple objects across vide o frames. However, current state-of-the-art methods mainly rely on a global opti mization technique and multi-stage cascade association strategy, and those appro aches often overlook the specific characteristics of assignment task in MOT and useful detection results that may represent occluded objects. To address these c hallenges, we propose a novel Track-Focused Online Multi-Object Tracker (TrackTr ack) with two key strategies: Track-Perspective-Based Association (TPA) and Trac k-Aware Initialization (TAI). The TPA strategy associates each track with the mo st suitable detection result by choosing the one with the minimum distance from all available detection results in a track-perspective manner. On the other hand , TAI precludes the generation of spurious tracks in the track-aware aspect by s uppressing track initialization of detection results that heavily overlap with c urrent active tracks and more confident detection results. Extensive experiments on MOT17, MOT20, and DanceTrack demonstrate that our TrackTrack outperforms cur rent state-of-the-art trackers, offering improved robustness and accuracy across diverse and challenging tracking scenarios.
****

Floxels: Fast Unsupervised Voxel Based Scene Flow Estimation
David T. Hoffmann, Syed Haseeb Raza, Hanqiu Jiang, Denis Tananaev, Steffen Kling enhoefer, Martin Meinke; Proceedings of the Computer Vision and Pattern Recognit ion Conference (CVPR), 2025, pp. 22328-22337
Scene flow estimation is a foundational task for many robotic applications, incl uding robust dynamic object detection, automatic labeling, and sensor synchroniz ation. Two types of approaches to the problem have evolved: 1) Supervised and 2) optimization-based methods. Supervised methods are fast during inference and ac hieve high-quality results, however, they are limited by the need for large amou nts of labeled training data and are susceptible to domain gaps. In contrast, un supervised test-time optimization methods do not face the problem of domain gaps but usually suffer from substantial runtime, exhibit artifacts, or fail to conv erge to the right solution. In this work, we mitigate several limitations of exi sting optimization-based methods. To this end, we 1) introduce a simple voxel gr id-based model that improves over the standard MLP-based formulation in multiple dimensions and 2) introduce a new multi-frame loss formulation. 3) We combine b oth contributions in our new method, termed Floxels. On the Argoverse 2 benchmar

k, Floxels is surpassed only by EulerFlow among unsupervised methods while achie
ving comparable performance at a fraction of the computational cost. Floxels ach
ieves a massive speedup of more than 60-140x over EulerFlow, reducing the runti
me from a day to 10 minutes per sequence. Over the faster but low-quality baseli
ne, NSFP, Floxels achieves a speedup of 14x.
****

LiveCC: Learning Video LLM with Streaming Speech Transcription at Scale
Joya Chen, Ziyun Zeng, Yiqi Lin, Wei Li, Zejun Ma, Mike Zheng Shou; Proceedings
of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2908
3-29095
Recent video large language models (Video LLMs) often depend on costly human ann
otations or proprietary APIs (e.g., GPT-4o) to produce training data, which limi
ts their training at scale. In this paper, we explore large-scale training for V
ideo LLM with cheap automatic speech recognition (ASR) transcripts. Specifically
, we propose a novel streaming training approach that densely interleaves the AS
R words and video frames according to their timestamps. Compared to previous stu
dies in vision-language representation with ASR, our method naturally fits the s
treaming characteristics of ASR, thus enabling the model to learn temporally-ali
gned, fine-grained vision-language modeling. To support the training algorithm,
we introduce a data pipeline for YouTube videos and their closed captions (CC),
resulting in \texttt Live-CC-10M pre-training set and \texttt Live-WhisperX-408
K high-quality supervised fine-tuning (SFT) set. Remarkably, even without SFT,
the pre-trained model \texttt LiveCC-7B demonstrates significant improvements i
n general video QA and exhibits a new capability in real-time video commentary.
To evaluate this, we carefully design a new benchmark \texttt LiveSports-3K , us
ing LLM-as-a-judge to measure the free-form commentary. Experiments show our fin
al model \texttt LiveCC-7B can surpass LLaVA-Video-72B in commentary quality ev
en working in a real-time mode. Meanwhile, it achieves state-of-the-art results
at the 7B scale on popular benchmarks such as VideoMME, demonstrating its broad
generalizability. All resources of this paper have been released at \href https:
//showlab.github.io/livecc showlab.github.io/livecc .
****

Identity-preserving Distillation Sampling by Fixed-Point Iterator
SeonHwa Kim, Jiwon Kim, Soobin Park, Donghoon Ahn, Jiwon Kang, Seungryong Kim, K
yong Hwan Jin, Eunju Cha; Proceedings of the Computer Vision and Pattern Recogni
tion Conference (CVPR), 2025, pp. 11115-11124
Score distillation sampling (SDS) demonstrates a powerful capability for text-co
nditioned 2D image and 3D object generation by distilling the knowledge from lea
rned score functions. However, SDS often suffers from blurriness caused by noisy
 gradients. When SDS meets the image editing, such degradations can be reduced b
y adjusting bias shifts using reference pairs, but the de-biasing techniques are
 still corrupted by erroneous gradients. To this end, we introduce Identity-pres
erving Distillation Sampling (IDS), which compensates for the gradient leading t
o undesired changes in the results. Based on the analysis that these errors come
 from the text-conditioned scores, a new regularization technique, called fixed-
point iterative regularization (FPR), is proposed to modify the score itself, dr
iving the preservation of the identity even including poses and structures. Than
ks to a self-correction by FPR, the proposed method provides clear and unambiguo
us representations corresponding to the given prompts in image-to-image editing
and editable neural radiance field (NeRF). The structural consistency between th
e source and the edited data is obviously maintained compared to other state-of-
the-art methods.
****

Progressive Focused Transformer for Single Image Super-Resolution
Wei Long, Xingyu Zhou, Leheng Zhang, Shuhang Gu; Proceedings of the Computer Vis
ion and Pattern Recognition Conference (CVPR), 2025, pp. 2279-2288
Transformer-based methods have achieved remarkable results in image super-resolu
tion tasks because they can capture non-local dependencies in low-quality input
images. However, this feature-intensive modeling approach is computationally exp
ensive because it calculates the similarities between numerous features that are

irrelevant to the query features when obtaining attention weights. These unnecessary similarity calculations not only degrade the reconstruction performance but also introduce significant computational overhead. How to accurately identify the features that are important to the current query features and avoid similarity calculations between irrelevant features remains an urgent problem. To address this issue, we propose a novel and effective Progressive Focused Transformer (PFT) that links all isolated attention maps in the network through Progressive Focused Attention (PFA) to focus attention on the most important tokens. PFA not only enables the network to capture more critical similar features, but also significantly reduces the computational cost of the overall network by filtering out irrelevant features before calculating similarities. Extensive experiments demonstrate the effectiveness of the proposed method, achieving state-of-the-art performance on various single image super-resolution benchmarks.
****

VladVA: Discriminative Fine-tuning of LVLMs
Yassine Ouali, Adrian Bulat, Alexandros Xenos, Anestis Zaganidis, Ioannis Maniadis Metaxas, Brais Martinez, Georgios Tzimiropoulos; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4101-4111
Contrastively-trained Vision-Language Models (VLMs) like CLIP have become the de facto approach for discriminative vision-language representation learning. However, these models have limited language understanding, often exhibiting a "bag of words" behavior. At the same time, Large Vision-Language Models (LVLMs), which combine vision encoders with LLMs, have been shown to be capable of detailed vision-language reasoning, yet their autoregressive nature renders them less suitable for discriminative tasks. In this work, we propose to combine "the best of both worlds": a new training approach for discriminative fine-tuning of LVLMs that results in strong discriminative and compositional capabilities. Essentially, our approach converts a generative LVLM into a discriminative one, unlocking its capability for powerful image-text discrimination combined with enhanced language understanding. Our contributions include (1) A carefully designed training/optimization framework that utilizes image-text pairs of variable length and granularity for training the model with both contrastive and next-token prediction losses. This is accompanied by ablation studies that justify the necessity of our framework's components. (2) A parameter-efficient adaptation method using a combination of soft prompting and LoRA adapters. (3) Significant improvements over state-of-the-art CLIP-like models of similar size, including standard image-text retrieval benchmarks and notable gains in compositionality.
****

FlexiDiT: Your Diffusion Transformer Can Easily Generate High-Quality Samples with Less Compute
Sotiris Anagnostidis, Gregor Bachmann, Yeongmin Kim, Jonas Kohler, Markos Georgopoulos, Artsiom Sanakoyeu, Yuming Du, Albert Pumarola, Ali Thabet, Edgar Schönfeld; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28316-28326
Despite their remarkable performance, modern Diffusion Transformers (DiTs) are hindered by substantial resource requirements during inference, stemming from the fixed and large amount of compute needed for each denoising step. In this work, we revisit the conventional static paradigm that allocates a fixed compute budget per denoising iteration and propose a dynamic strategy instead. Our simple and sample-efficient framework enables pre-trained DiT models to be converted into flexible ones --- dubbed FlexiDiT --- allowing them to process inputs at varying compute budgets. We demonstrate how a single flexible model can generate images without any drop in quality, while reducing the required FLOPs by more than 40% compared to their static counterparts, for both class-conditioned and text-conditioned image generation. Our method is general and agnostic to input and conditioning modalities. We show how our approach can be readily extended for video generation, where FlexiDiT models generate samples with up to 75% less compute without compromising performance.
****

WiLoR: End-to-end 3D Hand Localization and Reconstruction in-the-wild

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, Stefanos Zafeiriou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12242-12254

In recent years, 3D hand pose estimation methods have garnered significant attention due to their extensive applications in human-computer interaction, virtual reality, and robotics. In contrast, there has been a notable gap in hand detection pipelines, posing significant challenges in constructing effective real-world multi-hand reconstruction systems. In this work, we present a data-driven pipeline for efficient multi-hand reconstruction in the wild. The proposed pipeline is composed of two components: a real-time fully convolutional hand localization and a high-fidelity transformer-based 3D hand reconstruction model. To tackle the limitations of previous methods and build a robust and stable detection network, we introduce a large-scale dataset with over than 2M in-the-wild hand images with diverse lighting, illumination, and occlusion conditions. Our approach outperforms previous methods in both efficiency and accuracy on popular 2D and 3D benchmarks. Finally, we showcase the effectiveness of our pipeline to achieve smooth 3D hand tracking from monocular videos, without utilizing any temporal components. Code, models, and dataset are available at https://rolpotamias.github.io/WiLoR

****

HumanMM: Global Human Motion Recovery from Multi-shot Videos

Yuhong Zhang, Guanlin Wu, Ling-Hao Chen, Zhuokai Zhao, Jing Lin, Xiaoke Jiang, Jiamin Wu, Zhuoheng Li, Hao Frank Yang, Haoqian Wang, Lei Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1973-1983

In this paper, we present a novel framework designed to reconstruct long-sequence 3D human motion in the world coordinates from in-the-wild videos with multiple shot transitions. Such long-sequence in-the-wild motions are highly valuable to applications such as motion generation and motion understanding, but are of great challenge to be recovered due to abrupt shot transitions, partial occlusions, and dynamic backgrounds presented in such videos. Existing methods primarily focus on single-shot videos, where continuity is maintained within a single camera view, or simplify multi-shot alignment in camera space only. In this work, we tackle the challenges by integrating an enhanced camera pose estimation with Human Motion Recovery (HMR) by incorporating a shot transition detector and a robust alignment module for accurate pose and orientation continuity across shots. By leveraging a custom motion integrator, we effectively mitigate the problem of foot sliding and ensure temporal consistency in human pose. Extensive evaluations on our created multi-shot dataset from public 3D human datasets demonstrate the robustness of our method in reconstructing realistic human motion in world coordinates.

****

Removing Reflections from RAW Photos

Eric Kee, Adam Pikielny, Kevin Blackburn-Matzen, Marc Levoy; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 161-171

We describe a system to remove real-world reflections from images for consumer photography. Our system operates on linear (RAW) photos, and accepts an optional contextual photo looking in the opposite direction (e.g., the "selfie" camera on a mobile device). This optional photo disambiguates what should be considered the reflection. The system is trained solely on synthetic mixtures of real RAW photos, which we combine using a reflection simulation that is photometrically and geometrically accurate. Our system comprises a base model that accepts the captured photo and optional context photo as input, and runs at 256p, followed by an up-sampling model that transforms 256p images to full resolution. The system produces preview images at 1K in 4.5-6.5s on a MacBook or iPhone 14 Pro. We show SOTA results on RAW photos that were captured in the field to embody typical consumer photos, and show that training on RAW simulation data improves performance more than the architectural variations among prior works.

****

BiomedCoOp: Learning to Prompt for Biomedical Vision-Language Models

Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, Yiming Xiao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14766-14776

Recent advancements in vision-language models (VLMs), such as CLIP, have demonstrated substantial success in self-supervised representation learning for vision tasks. However, effectively adapting VLMs to downstream applications remains challenging, as their accuracy often depends on time-intensive and expertise-demanding prompt engineering, while full model fine-tuning is costly. This is particularly true for biomedical images, which, unlike natural images, typically suffer from limited annotated datasets, unintuitive image contrasts, and nuanced visual features. Recent prompt learning techniques, such as Context Optimization (CoOp) intend to tackle these issues, but still fall short in generalizability. Meanwhile, explorations in prompt learning for biomedical image analysis are still highly limited. In this work, we propose BiomedCoOp, a novel prompt learning framework that enables efficient adaptation of BiomedCLIP for accurate and highly generalizable few-shot biomedical image classification. Our approach achieves effective prompt context learning by leveraging semantic consistency with average prompt ensembles from Large Language Models (LLMs) and knowledge distillation with a statistics-based prompt selection strategy. We conducted comprehensive validation of our proposed framework on 11 medical datasets across 9 modalities and 10 organs against existing state-of-the-art methods, demonstrating significant improvements in both accuracy and generalizability. The code is publicly available at https://github.com/HealthX-Lab/BiomedCoOp.

****

AMR-Transformer: Enabling Efficient Long-range Interaction for Complex Neural Fluid Simulation

Zeyi Xu, Jinfan Liu, Kuangxu Chen, Ye Chen, Zhangli Hu, Bingbing Ni; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5804-5813

Accurately and efficiently simulating complex fluid dynamics is a challenging task that has traditionally relied on computationally intensive methods. Neural network-based approaches, such as convolutional and graph neural networks, have partially alleviated this burden by enabling efficient local feature extraction. However, they struggle to capture long-range dependencies due to limited receptive fields, and Transformer-based models, while providing global context, incur prohibitive computational costs. To tackle these challenges, we propose AMR-Transformer, an efficient and accurate neural CFD-solving pipeline that integrates a novel adaptive mesh refinement scheme with a Navier-Stokes constraint-aware fast pruning module. This design encourages long-range interactions between simulation cells and facilitates the modeling of global fluid wave patterns, such as turbulence and shockwaves. Experiments show that our approach achieves significant gains in efficiency while preserving critical details, making it suitable for high-resolution physical simulations with long-range dependencies. On CFDBench, PDEBench and a new shockwave dataset, our pipeline demonstrates up to an order-of-magnitude improvement in accuracy over baseline models. Additionally, compared to ViT, our approach achieves a reduction in FLOPs of up to 60 times.

****

MV-SSM: Multi-View State Space Modeling for 3D Human Pose Estimation

Aviral Chharia, Wenbo Gou, Haoye Dong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11590-11599

While significant progress has been made in single-view 3D human pose estimation, multi-view 3D human pose estimation remains challenging, particularly in terms of generalizing to new camera configurations. Existing attention-based transformers often struggle to accurately model the spatial arrangement of keypoints, especially in occluded scenarios. Additionally, they tend to overfit specific camera arrangements and visual scenes from training data, resulting in substantial performance drops in new settings. In this study, we introduce a novel Multi-View State Space Modeling framework, named MV-SSM, for robustly estimating 3D human keypoints. We explicitly model the joint spatial sequence at two distinct levels: the feature level from multi-view images and the person keypoint level. We pro

pose a Projective State Space (PSS) block to learn a generalized representation of joint spatial arrangements using state space modeling. Moreover, we modify Mamba's traditional scanning into an effective Grid Token-guided Bidirectional Scanning (GTBS), which is integral to the PSS block. Multiple experiments demonstrate that MV-SSM achieves strong generalization, outperforming state-of-the-art methods: +10.8 on AP25 on the challenging three-camera setting in CMU Panoptic, +7.0 on AP25 on varying camera arrangements, and +15.3 PCP on Campus A1 in cross-dataset evaluations. Project Website: https://aviralchharia.github.io/MV-SSM
****

HyperGLM: HyperGraph for Video Scene Graph Generation and Anticipation
Trong-Thuan Nguyen, Pha Nguyen, Jackson Cothren, Alper Yilmaz, Khoa Luu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29150-29160
Multimodal LLMs have advanced vision-language tasks but still struggle with understanding video scenes. To bridge this gap, Video Scene Graph Generation (VidSGG) has emerged to capture multi-object relationships across video frames. However, prior methods rely on pairwise connections, limiting their ability to handle complex multi-object interactions and reasoning. To this end, we propose Multimodal LLMs on a Scene HyperGraph (HyperGLM), promoting reasoning about multi-way interactions and higher-order relationships. Our approach uniquely integrates entity scene graphs, which capture spatial relationships between objects, with a procedural graph that models their causal transitions, forming a unified HyperGraph. Significantly, HyperGLM enables reasoning by injecting this unified HyperGraph into LLMs. Additionally, we introduce a new Video Scene Graph Reasoning (VSGR) dataset featuring 1.9M frames from third-person, egocentric, and drone views and supports five tasks: Scene Graph Generation, Scene Graph Anticipation, Video Question Answering, Video Captioning, and Relation Reasoning. Empirically, HyperGLM consistently outperforms state-of-the-art methods across five tasks, effectively modeling and reasoning complex relationships in diverse video scenes.
****

AnySat: One Earth Observation Model for Many Resolutions, Scales, and Modalities
Guillaume Astruc, Nicolas Gonthier, Clément Mallet, Loic Landrieu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19530-19540
Geospatial models must adapt to the diversity of Earth observation data in terms of resolutions, scales, and modalities. However, existing approaches expect fixed input configurations, which limits their practical applicability. We propose AnySat, a multimodal model based on joint embedding predictive architecture (JEPA) and scale-adaptive spatial encoders, allowing us to train a single model on highly heterogeneous data in a self-supervised manner. To demonstrate the advantages of this unified approach, we compile GeoPlex, a collection of 5 multimodal datasets with varying characteristics and 11 distinct sensors. We then train a single powerful model on these diverse datasets simultaneously. Once fine-tuned or probed, we achieve state-of-the-art results on the test sets of GeoPlex and for 6 external datasets across various environment monitoring tasks: land cover mapping, tree species identification, crop type classification, change detection, climate type classification, and segmentation of flood, burn scar, and deforestation. Our code and models are available at https://github.com/gastruc/AnySat.
****

FSFM: A Generalizable Face Security Foundation Model via Self-Supervised Facial Representation Learning
Gaojian Wang, Feng Lin, Tong Wu, Zhenguang Liu, Zhongjie Ba, Kui Ren; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24364-24376
This work asks: with abundant, unlabeled real faces, how to learn a robust and transferable facial representation that boosts various face security tasks with respect to generalization performance? We make the first attempt and propose a self-supervised pretraining framework to learn fundamental representations of real face images, FSFM, that leverages the synergy between masked image modeling (MIM) and instance discrimination (ID). We explore various facial masking strategie

s for MIM and present a simple yet powerful CRFR-P masking, which explicitly for ces the model to capture meaningful intra-region Consistency and challenging int er-region Coherency. Furthermore, we devise an ID network that naturally couples with MIM to establish underlying local-to-global Correspondence through tailore d self-distillation. These three learning objectives, namely 3C, empower encodin g both local features and global semantics of real faces. After pretraining, a v anilla ViT serves as a universal vision Foundation Model for downstream Face Sec urity tasks: cross-dataset deepfake detection, cross-domain face anti-spoofing, and unseen diffusion facial forgery detection. Extensive experiments on 10 publi c datasets demonstrate that our model transfers better than supervised pretraini ng, visual and facial self-supervised learning arts, and even outperforms task-s pecialized SOTA methods.

****

OVO-Bench: How Far is Your Video-LLMs from Real-World Online Video Understanding ?

Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, Pan Zhang, Yuhang Zang, Yuhang Cao , Conghui He, Jiaqi Wang; Proceedings of the Computer Vision and Pattern Recogni tion Conference (CVPR), 2025, pp. 18902-18913

Temporal Awareness, the ability to reason dynamically based on the timestamp whe n a question is raised, is the key distinction between offline and online video LLMs. Unlike offline models, which rely on complete videos for static, post hoc analysis, online models process video streams incrementally and dynamically adap t their responses based on the timestamp at which the question is posed. Despite its significance, temporal awareness has not been adequately evaluated in exist ing benchmarks. To fill this gap, we present OVO-Bench (Online-VideO-Benchmark), a novel video benchmark that emphasizes the importance of timestamps for advanc ed online video understanding capability benchmarking. OVO-Bench evaluates the a bility of video LLMs to reason and respond to events occurring at specific times tamps under three distinct scenarios: (1) Backward tracing: trace back to past e vents to answer the question. (2) Real-time understanding: understand and respon d to events as they unfold at the current timestamp. (3) Forward active respondi ng: delay the response until sufficient future information becomes available to answer the question accurately. OVO-Bench comprises 12 tasks, featuring 644 uniq ue videos and approximately human-curated 2,800 fine-grained meta-annotations wi th precise timestamps. We combine automated generation pipelines with human cura tion. With these high-quality samples, we further developed an evaluation pipeli ne to systematically query video LLMs along the video timeline. Evaluations of n ine Video-LLMs reveal that, despite advancements on traditional benchmarks, curr ent models struggle with online video understanding, showing a significant gap c ompared to human agents. We hope OVO-Bench will drive progress in video LLMs and inspire future research in online video reasoning. Our benchmark and code can b e accessed at https://joeleelyf.github.io/OVO-Bench.

****

AlignMamba: Enhancing Multimodal Mamba with Local and Global Cross-modal Alignme nt

Yan Li, Yifei Xing, Xiangyuan Lan, Xin Li, Haifeng Chen, Dongmei Jiang; Proceedi ngs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24774-24784

Cross-modal alignment is crucial for multimodal representation fusion due to the inherent heterogeneity between modalities. While Transformer-based methods have shown promising results in modeling inter-modal relationships, their quadratic computational complexity limits their applicability to long-sequence or large-sc ale data. Although recent Mamba-based approaches achieve linear complexity, thei r sequential scanning mechanism poses fundamental challenges in comprehensively modeling cross-modal relationships. To address this limitation, we propose Align Mamba, an efficient and effective method for multimodal fusion. Specifically, gr ounded in Optimal Transport, we introduce a local cross-modal alignment module t hat explicitly learns token-level correspondences between different modalities. Moreover, we propose a global cross-modal alignment loss based on Maximum Mean D

iscrepancy to implicitly enforce the consistency between different modal distrib
utions. Finally, the unimodal representations after local and global alignment a
re passed to the Mamba backbone for further cross-modal interaction and multimod
al fusion. Extensive experiments on complete and incomplete multimodal fusion ta
sks demonstrate the effectiveness and efficiency of the proposed method. For ins
tance, on the CMU-MOSI dataset, AlignMamba improves classification accuracy by 0
.9%, reduces GPU memory usage by 20.3%, and decreases inference time by 83.3%.
****

Blurry-Edges: Photon-Limited Depth Estimation from Defocused Boundaries
Wei Xu, Charles James Wagner, Junjie Luo, Qi Guo; Proceedings of the Computer Vi
sion and Pattern Recognition Conference (CVPR), 2025, pp. 432-441
Extracting depth information from photon-limited, defocused images is challengin
g because depth from defocus (DfD) relies on accurate estimation of defocus blur
, which is fundamentally sensitive to image noise. We present a novel approach t
o robustly measure object depths from photon-limited images along the defocused
boundaries. It is based on a new image patch representation, Blurry-Edges, that
explicitly stores and visualizes a rich set of low-level patch information, incl
uding boundaries, color, and smoothness. We develop a deep neural network archit
ecture that predicts the Blurry-Edges representation from a pair of differently
defocused images, from which depth can be calculated using a closed-form DfD rel
ation we derive. The experimental results on synthetic and real data show that o
ur method achieves the highest depth estimation accuracy on photon-limited image
s compared to a broad range of state-of-the-art DfD methods.
****

VideoComp: Advancing Fine-Grained Compositional and Temporal Alignment in Video-
Text Models
Dahun Kim, AJ Piergiovanni, Ganesh Mallya, Anelia Angelova; Proceedings of the C
omputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29060-29070
We introduce VideoComp, a benchmark and learning framework for advancing video-t
ext compositionality understanding, aimed at improving vision-language models (V
LMs) in fine-grained temporal alignment. Unlike existing benchmarks focused on s
tatic image-text compositionality or isolated single-event videos, our benchmark
 targets alignment in continuous multi-event videos. Leveraging video-text datas
ets with temporally localized event captions (e.g. ActivityNet-Captions, YouCook
2), we construct two compositional benchmarks, ActivityNet-Comp and YouCook2-Com
p. We create challenging negative samples with subtle temporal disruptions such
as reordering, action word replacement, partial captioning, and combined disrupt
ions. These benchmarks comprehensively test models' compositional sensitivity ac
ross extended, cohesive video-text sequences. To improve model performance, we p
ropose a hierarchical pairwise preference loss that strengthens alignment with t
emporally accurate pairs and gradually penalizes increasingly disrupted ones, en
couraging fine-grained compositional learning. To mitigate the limited availabil
ity of densely annotated video data, we introduce a pretraining strategy that co
ncatenates short video-caption pairs to simulate multi-event sequences. We evalu
ate video-text foundational models and large multimodal models (LMMs) on our ben
chmark, identifying both strengths and areas for improvement in compositionality
. Overall, our work provides a comprehensive framework for evaluating and enhanc
ing model capabilities in achieving fine-grained, temporally coherent video-text
 alignment.
****

One Model for ALL: Low-Level Task Interaction Is a Key to Task-Agnostic Image Fu
sion
Chunyang Cheng, Tianyang Xu, Zhenhua Feng, Xiaojun Wu, Zhangyong Tang, Hui Li, Z
eyang Zhang, Sara Atito, Muhammad Awais, Josef Kittler; Proceedings of the Compu
ter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28102-28112
Advanced image fusion methods mostly prioritise high-level missions, where task
interaction struggles with semantic gaps, requiring complex bridging mechanisms.
 In contrast, we propose to leverage low-level vision tasks from digital photogr
aphy fusion, allowing for effective feature interaction through pixel-level supe
rvision. This new paradigm provides strong guidance for unsupervised multimodal

fusion without relying on abstract semantics, enhancing task-shared feature learning for broader applicability. Owning to the hybrid image features and enhanced universal representations, the proposed GIFNet supports diverse fusion tasks, achieving high performance across both seen and unseen scenarios with a single model. Uniquely, experimental results reveal that our framework also supports single-modality enhancement, offering superior flexibility for practical applications. Our code will be available at https://github.com/AWCXV/GIFNet.

****

MICAS: Multi-grained In-Context Adaptive Sampling for 3D Point Cloud Processing

Feifei Shao, Ping Liu, Zhao Wang, Yawei Luo, Hongwei Wang, Jun Xiao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6616-6626

Point cloud processing (PCP) encompasses tasks like reconstruction, denoising, registration, and segmentation, each often requiring specialized models to address unique task characteristics. While in-context learning (ICL) has shown promise across tasks by using a single model with task-specific demonstration prompts, its application to PCP reveals significant limitations. We identify inter-task and intra-task sensitivity issues in current ICL methods for PCP, which we attribute to inflexible sampling strategies lacking context adaptation at the point and prompt levels. To address these challenges, we propose MICAS, an advanced ICL framework featuring a multi-grained adaptive sampling mechanism tailored for PCP. MICAS introduces two core components: task-adaptive point sampling, which leverages inter-task cues for point-level sampling, and query-specific prompt sampling, which selects optimal prompts per query to mitigate intra-task sensitivity. To our knowledge, this is the first approach to introduce adaptive sampling tailored to the unique requirements of point clouds within an ICL framework. Extensive experiments show that MICAS not only efficiently handles various PCP tasks but also significantly outperforms existing methods. Notably, it achieves a remarkable 4.1% improvement in the part segmentation task and delivers consistent gains across various PCP applications.

****

Can Text-to-Video Generation help Video-Language Alignment?

Luca Zanella, Massimiliano Mancini, Willi Menapace, Sergey Tulyakov, Yiming Wang, Elisa Ricci; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24097-24107

Recent video-language alignment models are trained on sets of videos, each with an associated positive caption and a negative caption generated by large language models. A problem with this procedure is that negative captions may introduce linguistic biases, i.e., concepts are seen only as negatives and never associated with a video. While a solution would be to collect videos for the negative captions, existing databases lack the fine-grained variations needed to cover all possible negatives. In this work, we study whether synthetic videos can help to overcome this issue. Our preliminary analysis with multiple generators shows that, while promising on some tasks, synthetic videos harm the performance of the model on others. We hypothesize this issue is linked to noise (semantic and visual) in the generated videos and develop a method, SynViTA, that accounts for those. SynViTA dynamically weights the contribution of each synthetic video based on how similar its target caption is w.r.t. the real counterpart. Moreover, a semantic consistency loss makes the model focus on fine-grained differences across captions, rather than differences in video appearance. Experiments show that, on average, SynViTA improves over existing methods on VideoCon test sets and SSv2-Temporal, SSv2-Events, and ATP-Hard benchmarks, being a first promising step for using synthetic videos when learning video-language models.

****

GoalFlow: Goal-Driven Flow Matching for Multimodal Trajectories Generation in End-to-End Autonomous Driving

Zebin Xing, Xingyu Zhang, Yang Hu, Bo Jiang, Tong He, Qian Zhang, Xiaoxiao Long, Wei Yin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1602-1611

We propose GoalFlow, an end-to-end autonomous driving method for generating high

-quality multimodal trajectories. In autonomous driving scenarios, there is rare ly a single suitable trajectory. Recent methods have increasingly focused on mod eling multimodal trajectory distributions. However, they suffer from trajectory selection complexity and reduced trajectory quality due to high trajectory diver gence and inconsistencies between guidance and scene information. To address the se issues, we introduce GoalFlow, a novel method that effectively constrains the generative process to produce high-quality, multimodal trajectories. To resolve the trajectory divergence problem inherent in diffusion-based methods, GoalFlow constrains the generated trajectories by introducing a goal point. GoalFlow est ablishes a novel scoring mechanism that selects the most appropriate goal point from the candidate points based on scene information. Furthermore, GoalFlow empl oys an efficient generative method, Flow Matching, to generate multimodal trajec tories, and incorporates a refined scoring mechanism to select the optimal traje ctory from the candidates. Our experimental results, validated on the Navsim, de monstrate that GoalFlow achieves state-of-the-art performance, delivering robust multimodal trajectories for autonomous driving. GoalFlow achieved PDMS of 90.3, significantly surpassing other methods. Compared with other diffusion-policy-ba sed methods, our approach requires only a single denoising step to obtain excell ent performance. The code is available at https://github.com/YvanYin/GoalFlow.

****

GuardSplat: Efficient and Robust Watermarking for 3D Gaussian Splatting

Zixuan Chen, Guangcong Wang, Jiahao Zhu, Jianhuang Lai, Xiaohua Xie; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 163 25-16335

3D Gaussian Splatting (3DGS) has recently created impressive 3D assets for vario us applications. However, considering security, capacity, invisibility, and trai ning efficiency, the copyright of 3DGS assets is not well protected as existing watermarking methods are unsuited for its rendering pipeline. In this paper, we propose GuardSplat, an innovative and efficient framework for watermarking 3DGS assets. Specifically, 1) We propose a CLIP-guided pipeline for optimizing the me ssage decoder with minimal costs. The key objective is to achieve high-accuracy extraction by leveraging CLIP's aligning capability and rich representations, de monstrating exceptional capacity and efficiency. 2) We tailor a Spherical-Harmon ic-aware (SH-aware) Message Embedding module for 3DGS, seamlessly embedding mess ages into the SH features of each 3D Gaussian while preserving the original 3D s tructure. This enables watermarking 3DGS assets with minimal fidelity trade-offs and prevents malicious users from removing the watermarks from the model files, meeting the demands for invisibility and security. 3) We present an Anti-distor tion Message Extraction module to improve robustness against various distortions . Experiments demonstrate that GuardSplat outperforms state-of-the-art and achie ves fast optimization speed.

****

Weakly Supervised Contrastive Adversarial Training for Learning Robust Features from Semi-supervised Data

Lilin Zhang, Chengpei Wu, Ning Yang; Proceedings of the Computer Vision and Patt ern Recognition Conference (CVPR), 2025, pp. 25718-25727

Existing adversarial training (AT) methods often suffer from incomplete perturba tion, meaning that not all non-robust features are perturbed when generating adv ersarial examples (AEs). This results in residual correlations between non-robus t features and labels, leading to suboptimal learning of robust features. Howeve r, achieving complete perturbation--perturbing as many non-robust features as po ssible--is challenging due to the difficulty in distinguishing robust and non-ro bust features and the sparsity of labeled data. To address these challenges, we propose a novel approach called Weakly Supervised Contrastive Adversarial Traini ng (WSCAT). WSCAT ensures complete perturbation for improved learning of robust features by disrupting correlations between non-robust features and labels throu gh complete AE generation over partially labeled data, grounded in information t heory. Extensive theoretical analysis and comprehensive experiments on widely ad opted benchmarks validate the superiority of WSCAT. Our code is available at htt ps://github.com/zhang-lilin/WSCAT.

****
From Poses to Identity: Training-Free Person Re-Identification via Feature Centralization
Chao Yuan, Guiwei Zhang, Changxiao Ma, Tianyi Zhang, Guanglin Niu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24409-24418
Person re-identification (ReID) aims to extract accurate identity representation features. However, during feature extraction, individual samples are inevitably affected by noise (background, occlusions, and model limitations). Considering that features from the same identity follow a normal distribution around identity centers after training, we propose a Training-Free Feature Centralization ReID framework (Pose2ID) by aggregating the same identity features to reduce individual noise and enhance the stability of identity representation, which preserves the feature's original distribution for following strategies such as re-ranking. Specifically, to obtain samples of the same identity, we introduce two components: Identity-Guided Pedestrian Generation: by leveraging identity features to guide the generation process, we obtain high-quality images with diverse poses, ensuring identity consistency even in complex scenarios such as infrared, and occlusion. Neighbor Feature Centralization: it explores each sample's potential positive samples from its neighborhood. Experiments demonstrate that our generative model exhibits strong generalization capabilities and maintains high identity consistency. With the Feature Centralization framework, we achieve impressive performance even with an ImageNet pre-trained model without ReID training, reaching mAP/Rank-1 of 52.81/78.92 on Market1501. Moreover, our method sets new state-of-the-art results across standard, cross-modality, and occluded ReID tasks, showcasing strong adaptability.
****
ColabSfM: Collaborative Structure-from-Motion by Point Cloud Registration
Johan Edstedt, André Mateus, Alberto Jaenal; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6573-6583
Structure-from-Motion (SfM) is the task of estimating 3D structure and camera poses from images. We define Collaborative SfM (ColabSfM) as sharing distributed SfM reconstructions. Sharing maps requires estimating a joint reference frame, which is typically referred to as registration. However, there is a lack of scalable methods and training datasets for registering SfM reconstructions. In this paper, we tackle this challenge by proposing the scalable task of point cloud registration for SfM reconstructions. We find that current registration methods cannot register SfM point clouds when trained on existing datasets. To this end, we propose a SfM registration dataset generation pipeline, leveraging partial reconstructions from synthetically generated camera trajectories for each scene. Finally, we propose a simple but impactful neural refiner on top of the SotA registration method RoITr that yields significant improvements, which we call RefineRoITr. Our extensive experimental evaluation shows that our proposed pipeline and model enables ColabSfM. Code is available at https://github.com/EricssonResearch/ColabSfM.
****
RoadSocial: A Diverse VideoQA Dataset and Benchmark for Road Event Understanding from Social Video Narratives
Chirag Parikh, Deepti Rawat, Rakshitha R. T., Tathagata Ghosh, Ravi Kiran Sarvadevabhatla; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19002-19011
We introduce RoadSocial, a large-scale, diverse VideoQA dataset tailored for generic road event understanding from social media narratives. Unlike existing datasets limited by regional bias, viewpoint bias and expert-driven annotations, RoadSocial captures the global complexity of road events with varied geographies, camera viewpoints (CCTV, handheld, drones) and rich social discourse. Our scalable semi-automatic annotation framework leverages Text LLMs and Video LLMs to generate comprehensive question-answer pairs across 12 challenging QA tasks, pushing the boundaries of road event understanding. RoadSocial is derived from social media videos spanning 14M frames and 414K social comments, resulting in a dataset

with 13.2K videos, 674 tags and 260K high-quality QA pairs. We evaluate 18 Video LLMs (open-source and proprietary, driving-specific and general-purpose) on our road event understanding benchmark. We also demonstrate RoadSocial's utility in improving road event understanding capabilities of general-purpose Video LLMs.
****

MangaNinja: Line Art Colorization with Precise Reference Following
Zhiheng Liu, Ka Leong Cheng, Xi Chen, Jie Xiao, Hao Ouyang, Kai Zhu, Yu Liu, Yujun Shen, Qifeng Chen, Ping Luo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5666-5677
Derived from diffusion models, MangaNinja specializes in the task of reference-guided line art colorization. We incorporate two thoughtful designs to ensure precise character detail transcription, including a patch shuffling module to facilitate correspondence learning between the reference color image and the target line art, and a point-driven control scheme to enable fine-grained color matching. Experiments on a self-collected benchmark demonstrate the superiority of our model over current solutions in terms of precise colorization. We further showcase the potential of the proposed interactive point control in handling challenging cases (*e.g.*, extreme poses and shadows), cross-character colorization, multi-reference harmonization, *etc.*, beyond the reach of existing algorithms.
****

Nonisotropic Gaussian Diffusion for Realistic 3D Human Motion Prediction
Cecilia Curreli, Dominik Muhle, Abhishek Saroha, Zhenzhang Ye, Riccardo Marin, Daniel Cremers; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1871-1882
Probabilistic human motion prediction aims to forecast multiple possible future movements from past observations. While current approaches report high diversity and realism, they often generate motions with undetected limb stretching and jitter. To address this, we introduce SkeletonDiffusion, a latent diffusion model that embeds an explicit inductive bias on the human body within its architecture and training. We present a nonisotropic Gaussian diffusion formulation that aligns with the natural kinematic structure of the human skeleton and models relationships between body parts. Results show that our approach outperforms isotropic alternatives, consistently generating realistic predictions while avoiding artifacts such as limb distortion. Additionally, we identify a limitation in commonly used diversity metrics, which may favor models that produce inconsistent limb lengths within the same sequence. SkeletonDiffusion sets a new benchmark on three real-world datasets, outperforming various baselines across multiple evaluation metrics. We release the code on our \href https://ceveloper.github.io/publications/skeletondiffusion/ project page .
****

Is Your World Simulator a Good Story Presenter? A Consecutive Events-Based Benchmark for Future Long Video Generation
Yiping Wang, Xuehai He, Kuan Wang, Luyao Ma, Jianwei Yang, Shuohang Wang, Simon Shaolei Du, Yelong Shen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13629-13638
The current state-of-the-art video generative models can produce commercial-grade videos with highly realistic details. However, they still struggle to coherently present multiple sequential events in specific short stories, which is foreseeable an essential capability for future long video generation scenarios. For example, top T2V generative models still fail to generate a video of the short simple story "how to put an elephant into a refrigerator." While existing detail-oriented benchmarks primarily focus on fine-grained metrics like aesthetic quality and spatial-temporal consistency, they fall short of evaluating models' abilities to handle event-level story presentation. To address this gap, we introduce StoryEval, a story-oriented benchmark specifically designed to assess text-to-video (T2V) models' story-completion capabilities. StoryEval features 423 prompts spanning 7 classes, each representing short stories composed of 2-4 consecutive events. We employ Vision-Language Models, such as GPT-4o and LLaVA-OV-Chat-72B, to verify the completion of each event in the generated videos, applying a unanimous voting method to enhance reliability. Our methods ensure high alignment with

human evaluations, and the evaluation of 11 models reveals its challenge, with none exceeding an average story-completion rate of 50%. StoryEval provides a new benchmark for advancing T2V models and highlights the challenges and opportunities in developing next-generation solutions for coherent story-driven video generation. Project website is available at https://ypwang61.github.io/project/StoryEval.

****

LookCloser: Frequency-aware Radiance Field for Tiny-Detail Scene
Xiaoyu Zhang, Weihong Pan, Chong Bao, Xiyu Zhang, Xiaojun Xiang, Hanqing Jiang, Hujun Bao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16122-16132

Humans perceive and comprehend their surroundings through information spanning multiple frequencies. In immersive scenes, people naturally scan their environment to grasp its overall structure while examining fine details of objects that capture their attention. However, current NeRF frameworks primarily focus on modeling either high-frequency local views or the broad structure of scenes with low-frequency information, limited to balance both. We introduce FA-NeRF, a novel frequency-aware framework for view synthesis that simultaneously captures the overall scene structure and high-definition details within a single NeRF model. To achieve this, we propose a 3D frequency quantification method that analyzes the scene's frequency distribution, enabling frequency-aware rendering. Our framework incorporates a frequency grid for fast convergence and querying, a frequency-aware feature re-weighting strategy to balance features across different frequency contents. Extensive experiments show that our method significantly outperforms existing approaches in modeling entire scenes while preserving fine details.

****

PICO: Reconstructing 3D People In Contact with Objects
Alpár Cseke, Shashank Tripathi, Sai Kumar Dwivedi, Arjun S. Lakshmipathy, Agniv Chatterjee, Michael J. Black, Dimitrios Tzionas; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1783-1794

Recovering 3D Human-Object Interaction (HOI) from single color images is challenging due to depth ambiguities, occlusions, and the huge variation in object shape and appearance. Thus, past work requires controlled settings such as known object shapes and contacts, and tackles only limited object classes. Instead, we need methods that generalize to natural images and novel object classes. We tackle this in two main ways:(1) We collect PICO-db, a new dataset of natural images uniquely paired with dense 3D contact correspondences on both body and object meshes. To this end, we use images from the recent DAMON dataset that are paired with annotated contacts, but only on a canonical 3D body. In contrast, we seek contact labels on both the body and the object. To infer these given an image, we retrieve an appropriate 3D object mesh from a database by leveraging vision foundation models. Then, we project DAMON's body contact patches onto the object via a novel method needing only 2 clicks per patch. This minimal human input establishes rich contact correspondences between bodies and objects. (2) We exploit our new dataset in a novel render-and-compare fitting method, called PICO-fit, to recover 3D body and object meshes in interaction. PICO-fit infers contact for the SMPL-X body, retrieves a likely 3D object mesh and contact from PICO-db for that object, and uses the contact to iteratively fit the 3D body and object meshes to image evidence via optimization. Uniquely, PICO-fit works well for many object categories that no existing method can tackle. This is crucial for scaling HOI understanding in the wild. Our data and code are available at https://pico.is.tue.mpg.de.

****

Convex Relaxation for Robust Vanishing Point Estimation in Manhattan World
Bangyan Liao, Zhenjun Zhao, Haoang Li, Yi Zhou, Yingping Zeng, Hao Li, Peidong Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15823-15832

Determining the vanishing points (VPs) in a Manhattan world, as a fundamental task in many 3D vision applications, consists of jointly inferring the line-VP association and locating each VP. Existing methods are, however, either sub-optimal

solvers or pursuing global optimality at a significant cost of computing time. In contrast to prior works, we introduce convex relaxation techniques to solve this task for the first time. Specifically, we employ a "soft" association scheme, realized via a truncated multi-selection error, that allows for joint estimation of VPs' locations and line-VP associations. This approach leads to a primal problem that can be reformulated into a quadratically constrained quadratic programming (QCQP) problem, which is then relaxed into a convex semidefinite programming (SDP) problem. To solve this SDP problem efficiently, we present a globally optimal outlier-robust iterative solver (called GlobustVP), which independently searches for one VP and its associated lines in each iteration, treating other lines as outliers. After each independent update of all VPs, the mutual orthogonality between the three VPs in a Manhattan world is reinforced via local refinement. Extensive experiments on both synthetic and real-world data demonstrate that GlobustVP achieves a favorable balance between efficiency, robustness, and global optimality compared to previous works. The code is publicly available at github.com/wu-cvgl/GlobustVP.

****

Linguistics-aware Masked Image Modeling for Self-supervised Scene Text Recognition

Yifei Zhang, Chang Liu, Jin Wei, Xiaomeng Yang, Yu Zhou, Can Ma, Xiangyang Ji; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9318-9328

Text images are unique in their dual nature, encompassing both visual and linguistic information. The visual component encompasses structural and appearance-based features, while the linguistic dimension incorporates contextual and semantic elements. In scenarios with degraded visual quality, linguistic patterns serve as crucial supplements for comprehension, highlighting the necessity of integrating both aspects for robust scene text recognition (STR). Contemporary STR approaches often use language models or semantic reasoning modules to capture linguistic features, typically requiring large-scale annotated datasets. Self-supervised learning, which lacks annotations, presents challenges in disentangling linguistic features related to the global context. Typically, sequence contrastive learning emphasizes the alignment of local features, while masked image modeling (MIM) tends to exploit local structures to reconstruct visual patterns, resulting in limited linguistic knowledge. In this paper, we propose a Linguistics-aware Masked Image Modeling (LMIM) approach, which channels the linguistic information into the decoding process of MIM through a separate branch. Specifically, we design a linguistics alignment module to extract vision-independent features as linguistic guidance using inputs with different visual appearances. As features extend beyond mere visual structures, LMIM must consider the global context to achieve reconstruction. Extensive experiments on various benchmarks quantitatively demonstrate our state-of-the-art performance, and attention visualizations qualitatively show the simultaneous capture of both visual and linguistic information. The code is available at https://github.com/zhangyifei01/LMIM.

****

FruitNinja: 3D Object Interior Texture Generation with Gaussian Splatting

Fangyu Wu, Yuhao Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11051-11060

In the real world, objects reveal internal textures when sliced or cut, yet this behavior is not well-studied in 3D generation tasks today. For example, slicing a virtual 3D watermelon should reveal flesh and seeds. Given that no available dataset captures an object's full internal structure and collecting data from all slices is impractical, generative methods become the obvious approach. However, current 3D generation and inpainting methods often focus on visible appearance and overlook internal textures. To bridge this gap, we introduce FruitNinja, the first method to generate internal textures for 3D objects undergoing geometric and topological changes. Our approach produces objects via 3D Gaussian Splatting (3DGS) with both surface and interior textures synthesized, enabling real-time slicing and rendering without additional optimization. FruitNinja leverages a pre-trained diffusion model to progressively inpaint cross-sectional views and ap

plies voxel-grid-based smoothing to achieve cohesive textures throughout the obj
ect. Our OpaqueAtom GS strategy overcomes 3DGS limitations by employing densely
distributed opaque Gaussians, avoiding biases toward larger particles that desta
bilize training and sharp color transitions for fine-grained textures. Experime
ntal results show that FruitNinja substantially outperforms existing approaches,
 showcasing unmatched visual quality in real-time rendered internal views across
 arbitrary geometry manipulations. Project page: https://fanguw.github.io/FruitN
inja3D.
****

Scaling up Image Segmentation across Data and Tasks
Pei Wang, Zhaowei Cai, Hao Yang, Ashwin Swaminathan, R. Manmatha, Stefano Soatto
; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR),
2025, pp. 4573-4583
Traditional segmentation models, while effective in isolated tasks, often fail t
o generalize to more complex and open-ended segmentation problems, such as free-
form, open-vocabulary, and in-the-wild scenarios. To bridge this gap, we propose
 to scale up image segmentation across diverse datasets and tasks such that the
knowledge across different tasks and datasets can be integrated while improving
the generalization ability. QueryMeldNet, a novel segmentation framework, is int
roduced and designed to scale seamlessly across both data size and task diversit
y. It is built upon a dynamic object query mechanism called query meld, which fu
ses different types of queries using cross-attention. This hybrid approach enabl
es the model to balance between instance- and stuff-level segmentation, providin
g enhanced scalability for handling diverse object types. We further enhance sca
lability by leveraging synthetic data-generating segmentation masks and captions
 for pixel-level and open-vocabulary tasks-drastically reducing the need for cos
tly human annotations. By training on multiple datasets and tasks at scale, Quer
yMeldNet continuously improves performance as the volume and diversity of data a
nd tasks increase. It exhibits strong generalization capabilities, boosting perf
ormance in open-set segmentation tasks SeginW by 7 points. These advancements ma
rk a key step toward universal, scalable segmentation models capable of addressi
ng the demands of real-world applications.
****

Take the Bull by the Horns: Learning to Segment Hard Samples
Yuan Guo, Jingyu Kong, Yu Wang, Yuping Duan; Proceedings of the Computer Vision
and Pattern Recognition Conference (CVPR), 2025, pp. 15642-15652
Medical image segmentation is vital for clinical applications, with hard samples
 playing a key role in segmentation accuracy. We propose an effective image segm
entation framework that includes mechanisms for identifying and segmenting hard
samples. It derives a novel image segmentation paradigm: 1) Learning to identify
 hard samples: automatically selecting inherent hard samples from different data
sets, and 2) Learning to segment hard samples: achieving the segmentation of har
d samples through effective feature augmentation on dedicated networks. We name
our method `Learning to Segment hard samples' (L2S). The hard sample identificat
ion module comprises a backbone model and a classifier, which dynamically uncove
rs inherent dataset patterns. The hard sample segmentation module utilizes the d
iffusion process for feature augmentation and incorporates a more sophisticated
segmentation network to achieve precise segmentation. We justify our motivation
through solid theoretical analysis and extensive experiments. Evaluations across
 various modalities show that our L2S outperforms other SOTA methods, particular
ly by substantially improving the segmentation accuracy of hard samples. On ISIC
 dataset, our L2S improves the Dice score on hard samples and overall segmentati
on by 8.97% and 1.01%, respectively, compared to SOTA methods. The code is avail
able at \href https://github.com/TqlYuanGie/L2S https://github.com/TqlYuanGie/L
2S.
****

MIMO: A Medical Vision Language Model with Visual Referring Multimodal Input and
 Pixel Grounding Multimodal Output
Yanyuan Chen, Dexuan Xu, Yu Huang, Songkun Zhan, Hanpin Wang, Dongxue Chen, Xuep
ing Wang, Meikang Qiu, Hang Li; Proceedings of the Computer Vision and Pattern R

ecognition Conference (CVPR), 2025, pp. 24732-24741
Currently, medical vision language models are widely used in medical vision ques tion answering tasks. However, existing models are confronted with two issues: f or input, the model only relies on text instructions and lacks direct understand ing of visual clues in the image; for output, the model only gives text answers and lacks connection with key areas in the image. To address these issues, we pr opose a unified medical vision language model MIMO, with visual referring Multim odal Input and pixel grounding Multimodal Output. MIMO can not only combine visu al clues and textual instructions to understand complex medical images and seman tics, but can also ground medical terminologies in textual output within the ima ge. To overcome the scarcity of relevant data in the medical field, we propose M IMOSeg, a comprehensive medical multimodal dataset including 895K samples. MIMOS eg is constructed from four different perspectives, covering basic instruction f ollowing and complex question answering with multimodal input and multimodal out put. We conduct experiments on several downstream medical multimodal tasks. Exte nsive experimental results verify that MIMO can uniquely combine visual referrin g and pixel grounding capabilities, which are not available in previous models. Our project can be found in https://github.com/pkusixspace/MIMO.
****

Bias for Action: Video Implicit Neural Representations with Bias Modulation
Alper Kayabasi, Anil Kumar Vadathya, Guha Balakrishnan, Vishwanath Saragadam; Pr oceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025 , pp. 27999-28008
We propose a new continuous video modeling framework based on implicit neural re presentations (INRs) called ActINR. At the core of our approach is the observati on that INRs can be considered as a learnable dictionary, with the shapes of the basis functions governed by the weights of the INR, and their locations governe d by the biases. Given compact non-linear activation functions, we hypothesize t hat an INR's biases are suitable to capture motion across images, and facilitate compact representations for video sequences. Using these observations, we desig n ActINR to share INR weights across frames of a video sequence, while using uni que biases for each frame. We further model the biases as the output of a separa te INR conditioned on time index to promote smoothness. By training the video IN R and this bias INR together, we demonstrate unique capabilities, including 10x video slow motion, 4x spatial super resolution along with 2x slow motion, denois ing, and video inpainting. ActINR performs remarkably well across numerous video processing tasks (often achieving more than 6dB improvement), setting a new sta ndard for continuous modeling of videos.
****

Bridging Past and Future: End-to-End Autonomous Driving with Historical Predicti on and Planning
Bozhou Zhang, Nan Song, Xin Jin, Li Zhang; Proceedings of the Computer Vision an d Pattern Recognition Conference (CVPR), 2025, pp. 6854-6863
End-to-end autonomous driving unifies tasks in a differentiable framework, enabl ing planning-oriented optimization and attracting growing attention.Current meth ods aggregate historical information either through dense historical bird's-eye-view (BEV) features or by querying a sparse memory bank, following paradigms inh erited from detection.However, we argue that these paradigms either omit histori cal information in motion planning or fail to align with its multi-step nature, which requires predicting or planning multiple future time steps. In line with t he philosophy of "future is a continuation of past", we propose **BridgeAD**, wh ich reformulates motion and planning queries as multi-step queries to differenti ate the queries for each future time step. This design enables the effective use of historical prediction and planning by applying them to the appropriate parts of the end-to-end system based on the time steps, which improves both perceptio n and motion planning. Specifically, historical queries for the current frame ar e combined with perception, while queries for future frames are integrated with motion planning. In this way, we bridge the gap between past and future by aggre gating historical insights at every time step, enhancing the overall coherence a nd accuracy of the end-to-end autonomous driving pipeline. Extensive experiments

on the nuScenes dataset in both open-loop and closed-loop settings demonstrate that BridgeAD achieves state-of-the-art performance. We will make our code and models publicly available.
****

Blood Flow Speed Estimation with Optical Coherence Tomography Angiography Images
Wensheng Cheng, Zhenghong Li, Jiaxiang Ren, Hyomin Jeong, Congwu Du, Yingtian Pan, Haibin Ling; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10466-10475
Estimating blood flow speed is essential in many medical and physiological applications, yet it is extremely challenging due to complex vascular structure and flow dynamics, particularly for cerebral cortex regions. Existing techniques, such as Optical Doppler Tomography (ODT), generally require complex hardware control and signal processing, and still suffer from inherent system-level artifacts. To address these challenges, we propose a new learning-based approach named OCTA-Flow, which directly estimates vascular blood flow speed from Optical Coherence Tomography Angiography (OCTA) images that are commonly used for vascular structure analysis. OCTA-Flow employs several novel components to achieve this goal. First, using an encoder-decoder architecture, OCTA-Flow leverages ODT data as pseudo labels during training, thus bypassing the difficulty of collecting ground truth data. Second, to capture the relationship between vessels of varying scales and their flow speed, we design an Adaptive Window Fusion module that employs multiscale window attention. Third, to mitigate ODT artifacts, we incorporate a Conditional Random Field Decoder that promotes smoothness and consistency in the estimated blood flow. Together, these innovations enable OCTA-Flow to effectively produce accurate flow estimation, suppress the artifacts in ODT, and enhance practicality, benefiting from the established techniques of OCTA data acquisition. The code and data are available at https://github.com/Spritea/OCTA-Flow.
****

DreamTrack: Dreaming the Future for Multimodal Visual Object Tracking
Mingzhe Guo, Weiping Tan, Wenyu Ran, Liping Jing, Zhipeng Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7201-7210
Aiming to achieve class-agnostic perception in visual object tracking, current trackers commonly formulate tracking as a one-shot detection problem with the template-matching architecture. Despite the success, severe environmental variations in long-term tracking raise challenges to generalizing the tracker in novel situations. Temporal trackers try to fix it by preserving the time-validity of target information with historical predictions, e.g., updating the template. However, solely transmitting the previous observations instead of learning from them leads to an inferior capability of understanding the tracking scenario from past experience, which is critical for the generalization in new frames. To address this issue, we reformulate temporal learning in visual tracking as a History-to-Future process and propose a novel tracking framework DreamTrack. Our DreamTrack learns the temporal dynamics from past observations to dream the future variations of the environment, which boosts the generalization with the extended future information from history. Considering the uncertainty of future variation, multimodal prediction is designed to infer the target trajectory of each possible future situation. The experiments demonstrate that our DreamTrack achieves leading performance with real-time inference speed. In particular, DreamTrack obtains SUC scores of 76.6%/87.9% on LaSOT/TrackingNet, surpassing all recent SOTA trackers.
****

OmniStyle: Filtering High Quality Style Transfer Data at Scale
Ye Wang, Ruiqi Liu, Jiang Lin, Fei Liu, Zili Yi, Yilin Wang, Rui Ma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7847-7856
In this paper, we introduce OmniStyle-1M, a large-scale paired style transfer dataset comprising over one million content-style-stylized image triplets across 1,000 diverse style categories, each enhanced with textual descriptions and instruction prompts. We show that OmniStyle-1M can not only improve the generalizatio

n of style transfer models through supervised training but also facilitates prec
ise control over target stylization. Especially, to ensure the quality of the da
taset, we introduce OmniFilter, a comprehensive style transfer quality assessmen
t framework, which filters high-quality triplets based on content preservation,
style consistency, and aesthetic appeal. Building upon this foundation, we propo
se OmniStyle, a framework based on the Diffusion Transformer (DiT) architecture
designed for high-quality and efficient style transfer. This framework supports
both instruction-guided and image-guided style transfer, generating high resolut
ion outputs with exceptional detail. Extensive qualitative and quantitative eval
uations demonstrate OmniStyle's superior performance compared to existing approa
ches, highlighting its efficiency and versatility. OmniStyle-1M and its accompan
ying methodologies provide a significant contribution to advancing high-quality
style transfer, offering a valuable resource for the research community.
****

EIDT-V: Exploiting Intersections in Diffusion Trajectories for Model-Agnostic, Z
ero-Shot, Training-Free Text-to-Video Generation
Diljeet Jagpal, Xi Chen, Vinay P. Namboodiri; Proceedings of the Computer Vision
 and Pattern Recognition Conference (CVPR), 2025, pp. 18219-18228
Zero-shot, training-free, image-based text-to-video generation is an emerging ar
ea that aims to generate videos using existing image-based diffusion models. Cur
rent methods in this space require specific architectural changes to image-gener
ation models, which limit their adaptability and scalability.In contrast to such
 methods, we provide a model-agnostic approach. We use intersections in diffusio
n trajectories, working only with the latent values. We could not obtain localiz
ed frame-wise coherence and diversity using only the intersection of trajectorie
s. Thus, we instead use a grid-based approach. An in-context trained LLM is used
 to generate coherent frame-wise prompts; another is used to identify difference
s between frames. Based on these, we obtain a CLIP-based attention mask that con
trols the timing of switching the prompts for each grid cell. Earlier switching
results in higher variance, while later switching results in more coherence. The
refore, Our approach can ensure appropriate control between coherence and varian
ce for the frames.Our approach results in state-of-the-art performance while bei
ng more flexible when working with diverse image-generation models. The empirica
l analysis using quantitative metrics and user studies confirms our model's supe
rior temporal consistency, visual fidelity and user satisfaction, thus providing
 a novel way to obtain training-free, image-based text-to-video generation.
****

Cross-View Completion Models are Zero-shot Correspondence Estimators
Honggyu An, Jin Hyeon Kim, Seonghoon Park, Jaewoo Jung, Jisang Han, Sunghwan Hon
g, Seungryong Kim; Proceedings of the Computer Vision and Pattern Recognition Co
nference (CVPR), 2025, pp. 1103-1115
In this work, we analyze new aspects of cross-view completion, mainly through th
e analogy of cross-view completion and traditional self-supervised correspondenc
e learning algorithms. Based on our analysis, we reveal that the cross-attention
 map of Croco-v2, best reflects this correspondence information compared to othe
r correlations from the encoder or decoder features. We further verify the effec
tiveness of the cross-attention map by evaluating on both zero-shot and supervis
ed dense geometric correspondence and multi-frame depth estimation.
****

Multi-party Collaborative Attention Control for Image Customization
Han Yang, Chuanguang Yang, Qiuli Wang, Zhulin An, Weilun Feng, Libo Huang, Yongj
un Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CV
PR), 2025, pp. 7942-7951
The rapid development of diffusion models has fueled a growing demand for custom
ized image generation. However, current customization methods face several limit
ations: 1) typically accept either image or text conditions alone; 2) customizat
ion in complex visual scenarios often leads to subject leakage or confusion; 3)
image-conditioned outputs tend to suffer from inconsistent backgrounds; and 4) h
igh computational costs. To address these issues, this paper introduces Multi-pa
rty Collaborative Attention Control (MCA-Ctrl), a tuning-free approach that enab

les high-quality image customization under both text and complex visual conditio ns. Specifically, MCA-Ctrl leverages two key operations within the self-attentio n layer to coordinate multiple parallel diffusion processes and guide the target image generation. This approach allows MCA-Ctrl to capture the content and appe arance of specific subjects while maintaining semantic consistency with the cond itional input. Additionally, to mitigate subject leakage and confusion issues co mmon in complex visual scenarios, we introduce a Subject Localization Module tha t extracts precise subject and editable image layers based on user instructions. Extensive quantitative and human evaluation experiments demonstrate that MCA-Ct rl outperforms previous methods in zero-shot image customization, effectively ad dressing the aforementioned issues.

****

Reproducible Vision-Language Models Meet Concepts Out of Pre-Training

Ziliang Chen, Xin Huang, Xiaoxuan Fan, Keze Wang, Yuyu Zhou, Quanlong Guan, Lian g Lin; Proceedings of the Computer Vision and Pattern Recognition Conference (CV PR), 2025, pp. 14701-14711

Contrastive Language-Image Pre-training (CLIP) models as a milestone of modern m ultimodal intelligence, its generalization mechanism grasped massive research in terests in the community. While existing studies limited in the scope of pre-tra ining knowledge, hardly underpinned its generalization to countless open-world c oncepts absent from the pre-training regime. This paper dives into such Out-of-P re-training (OOP) generalization problem from a holistic perspective. We propose LAION-Beyond benchmark to isolate the evaluation of OOP concepts from pre-train ing knowledge, with regards to OpenCLIP and its reproducible variants derived fr om LAION datasets. Empirical analysis evidences that despite image features of O OP concepts born with significant category margins, their zero-shot transfer sig nificantly fails due to the poor image-text alignment. To this, we elaborate the "name-tuning" methodology with its theoretical merits in terms of OOP generaliz ation, then propose few-shot name learning (FSNL) and zero-shot name learning (Z SNL) algorithms to achieve OOP generalization in a data-efficient manner. Their superiority have been further verified in our comprehensive experiments.

****

Through-The-Mask: Mask-based Motion Trajectories for Image-to-Video Generation

Guy Yariv, Yuval Kirstain, Amit Zohar, Shelly Sheynin, Yaniv Taigman, Yossi Adi, Sagie Benaim, Adam Polyak; Proceedings of the Computer Vision and Pattern Recog nition Conference (CVPR), 2025, pp. 18198-18208

We consider the task of Image-to-Video (I2V) generation, which involves transfor ming static images into realistic video sequences based on a textual description . While recent advancements produce photorealistic outputs, they frequently stru ggle to create videos with accurate and consistent object motion, especially in multi-object scenarios. To address these limitations, we propose a two-stage com positional framework that decomposes I2V generation into: (i) An explicit interm ediate representation generation stage, followed by (ii) A video generation stag e that is conditioned on this representation. Our key innovation is the introduc tion of a mask-based motion trajectory as an intermediate representation, that c aptures both semantic object information and motion, enabling an expressive but compact representation of motion and semantics. To incorporate the learned repre sentation in the second stage, we utilize object-level attention objectives. Spe cifically, we consider a spatial, per-object, masked-cross attention objective, integrating object-specific prompts into corresponding latent space regions and a masked spatio-temporal self-attention objective, ensuring frame-to-frame consi stency for each object. We evaluate our method on challenging benchmarks with mu lti-object and high-motion scenarios and empirically demonstrate that the propos ed method achieves state-of-the-art results in temporal coherence, motion realis m, and text-prompt faithfulness. Additionally, we introduce SA-V-128, a new chal lenging benchmark for single-object and multi-object I2V generation, and demonst rate our method's superiority on this benchmark. Project page is available at ht tps://guyyariv.github.io/TTM/

****

MAGE : Single Image to Material-Aware 3D via the Multi-View G-Buffer Estimation

Model
Haoyuan Wang, Zhenwei Wang, Xiaoxiao Long, Cheng Lin, Gerhard Hancke, Rynson W.H. Lau; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10985-10995
With advances in deep learning models and the availability of large-scale 3D datasets, we have recently witnessed significant progress in single-view 3D reconstruction. However, existing methods often fail to reconstruct physically based material properties given a single image, limiting their applicability in complicated scenarios. This paper presents a novel approach (named MAGE) for generating 3D geometry with realistic decomposed material properties given a single image as input. Our method leverages inspiration from traditional computer graphics deferred rendering pipelines to introduce a multi-view G-buffer estimation model. The proposed model estimates G-buffers for various views as multi-domain images, including XYZ coordinates, normals, albedo, roughness, and metallic properties from a single-view RGB image. To address the inherent ambiguity and inconsistency in generating G-buffers simultaneously, we also formulate a deterministic network from the pretrained diffusion models and propose a lighting response loss that enforces consistency across these domains using PBR principles. Finally, we propose a large-scale synthetic dataset rich in material diversity for our model training. Experimental results demonstrate the effectiveness of our method in producing high-quality 3D meshes with rich material properties. Our code and dataset can be found at https://www.whyy.site/paper/mage.
****
Segment Anything, Even Occluded
Wei-En Tai, Yu-Lin Shih, Cheng Sun, Yu-Chiang Frank Wang, Hwann-Tzong Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29385-29394
Amodal instance segmentation, which aims to detect and segment both visible and invisible parts of objects in images, plays a crucial role in various applications, including autonomous driving, robotic manipulation, and scene understanding. While existing methods require training both front-end detectors and mask decoders jointly, this approach lacks flexibility and fails to leverage the strengths of pre-existing modal detectors. To address this limitation, we propose SAMEO, a novel framework that adapts the Segment Anything Model (SAM) as a versatile mask decoder capable of interfacing with various front-end detectors to enable mask prediction even for partially occluded objects. Acknowledging the constraints of limited amodal segmentation datasets, we introduce Amodal-LVIS, a large-scale synthetic dataset comprising 300K images derived from the modal LVIS and LVVIS datasets. This dataset significantly expands the training data available for amodal segmentation research. Our experimental results demonstrate that our approach, when trained on the newly extended dataset, including Amodal-LVIS, achieves remarkable zero-shot performance on both COCOA-cls and D2SA benchmarks, highlighting its potential for generalization to unseen scenarios.
****
HOT3D: Hand and Object Tracking in 3D from Egocentric Multi-View Videos
Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, Tomas Hodan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7061-7071
We introduce HOT3D, a publicly available dataset for egocentric hand and object tracking in 3D. The dataset offers over 833 minutes (3.7M+ images) of recordings that feature 19 subjects interacting with 33 diverse rigid objects. In addition to simple pick-up, observe, and put-down actions, the subjects perform actions typical for a kitchen, office, and living room environment. The recordings include multiple synchronized data streams containing egocentric multi-view RGB/monochrome images, eye gaze signal, scene point clouds, and 3D poses of cameras, hands, and objects. The dataset is recorded with two headsets from Meta: Project Aria, which is a research prototype of AI glasses, and Quest 3, a virtual-reality headset that has shipped millions of units. Ground-truth poses were obtained by a motion-capture system using small optical markers attached to hands and objects

. Hand annotations are provided in the UmeTrack and MANO formats, and objects ar e represented by 3D meshes with PBR materials obtained by an in-house scanner. I n our experiments, we demonstrate the effectiveness of multi-view egocentric dat a for three popular tasks: 3D hand tracking, model-based 6DoF object pose estima tion, and 3D lifting of unknown in-hand objects. The evaluated multi-view method s, whose benchmarking is uniquely enabled by HOT3D, significantly outperform the ir single-view counterparts.
****

DELT: A Simple Diversity-driven EarlyLate Training for Dataset Distillation
Zhiqiang Shen, Ammar Sherif, Zeyuan Yin, Shitong Shao; Proceedings of the Comput er Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4797-4806
Recent advances in dataset distillation have led to solutions in two main direct ions. The conventional batch-to-batch matching mechanism is ideal for small-scal e datasets and includes bi-level optimization methods on models and syntheses, s uch as FRePo, RCIG, and RaT-BPTT, as well as other methods like distribution mat ching, gradient matching, and weight trajectory matching. Conversely, batch-to-g lobal matching typifies decoupled methods, which are particularly advantageous f or large-scale datasets. This approach has garnered substantial interest within the community, as seen in SRe^2L, G-VBSM, WMDD, and CDA. A primary challenge wit h the second approach is the lack of diversity among syntheses within each class since samples are optimized independently and the same global supervision signa ls are reused across different synthetic images. In this study, we propose a new Diversity-driven EarlyLate Training (DELT) scheme to enhance the diversity of i mages in batch-to-global matching with less computation. Our approach is concept ually simple yet effective, it partitions predefined IPC samples into smaller su btasks and employs local optimizations to distill each subset into distributions from distinct phases, reducing the uniformity induced by the unified optimizati on process. These distilled images from the subtasks demonstrate effective gener alization when applied to the entire task. We conduct extensive experiments on C IFAR, Tiny-ImageNet, ImageNet-1K, and its sub-datasets. Our approach outperforms the previous state-of-the-art by 2~5% on average across different datasets and IPCs (images per class), increasing diversity per class by more than 5% while re ducing synthesis time by up to 39.3% for enhancing the training efficiency.
****

MESC-3D:Mining Effective Semantic Cues for 3D Reconstruction from a Single Image
Shaoming Li, Qing Cai, Songqi Kong, Runqing Tan, Heng Tong, Shiji Qiu, Yongguo J iang, Zhi Liu; Proceedings of the Computer Vision and Pattern Recognition Confer ence (CVPR), 2025, pp. 16912-16921
Reconstructing 3D shapes from a single image plays an important role in computer vision. Many methods have been proposed and achieve impressive performance. How ever, existing methods mainly focus on extracting semantic information from imag es and then simply concatenating it with 3D point clouds without further explori ng the concatenated semantics. As a result, these entangled semantic features si gnificantly hinder the reconstruction performance. In this paper, we propose a n ovel single-image 3D reconstruction method called Mining Effective Semantic Cues for 3D Reconstruction from a Single Image (MESC-3D), which can actively mine ef fective semantic cues from entangled features. Specifically, we design an Effect ive Semantic Mining Module, which incorporates a point-selection map establish c onnections between point clouds and semantic attributes, enabling the point clou ds to autonomously select the necessary information. Furthermore, to address the potential insufficiencies in semantic information from a single image, such as occlusions, inspired by the human ability to represent 3D objects using prior kn owledge drawn from daily experiences, we introduce a 3D Semantic Prior Learning Module. This module employs contrastive learning paradigm along with a learnable text prompt to incorporates semantic understanding of spatial structures, enabl ing the model to interpret and reconstruct 3D objects with greater accuracy and realism, closely mirroring human perception of complex 3D environments. Extensiv e evaluations show that our method achieves significant improvements in reconstr uction quality and robustness compared to prior works. Additionally, further exp eriments validate the strong generalization capabilities and excels in zero-shot

preformance on unseen classes. Code is available at https://github.com/QINGQINGLE/MESC-3D.

****

RoboBrain: A Unified Brain Model for Robotic Manipulation from Abstract to Concrete

Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, Xinda Xue, Qinghang Su, Huaihai Lyu, Xiaolong Zheng, Jiaming Liu, Zhongyuan Wang, Shanghang Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1724-1734

Recent advancements in Multimodal Large Language Models (MLLMs) have shown remarkable capabilities across various multimodal contexts. However, their application in robotic scenarios, particularly for long-horizon manipulation tasks, reveals significant limitations. These limitations arise from the current MLLMs lacking three essential robotic brain capabilities: Planning Capability, which involves decomposing complex manipulation instructions into manageable sub-tasks; Affordance Perception, the ability to recognize and interpret the affordances of interactive objects; and Trajectory Prediction, the foresight to anticipate the complete manipulation trajectory necessary for successful execution. To enhance the robotic brain's core capabilities from abstract to concrete, we introduce ShareRobot, a high-quality heterogeneous dataset that labels multi-dimensional information such as task planning, object affordance, and end-effector trajectory. ShareRobot's diversity and accuracy have been meticulously refined by three human annotators. Building on this dataset, we developed RoboBrain, an MLLM-based model that combines robotic and general multi-modal data, utilizes a multi-stage training strategy, and incorporates long videos and high-resolution images to improve its robotic manipulation capabilities.Extensive experiments demonstrate that RoboBrain achieves state-of-the-art performance across various obotic tasks, highlighting its potential to advance robotic brain capabilities.

****

Advancing Multiple Instance Learning with Continual Learning for Whole Slide Imaging

Xianrui Li, Yufei Cui, Jun Li, Antoni B. Chan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20800-20809

Advances in medical imaging and deep learning have propelled progress in whole slide image (WSI) analysis, with multiple instance learning (MIL) showing promise for efficient and accurate diagnostics. However, conventional MIL models often lack adaptability to evolving datasets, as they rely on static training that cannot incorporate new information without extensive retraining. Applying continual learning (CL) to MIL models is a possible solution, but often sees limited improvements. In this paper, we analyze CL in the context of attention MIL models and find that the model forgetting is mainly concentrated in the attention layers of the MIL model. Using the results of this analysis we propose two components for improving CL on MIL:Attention Knowledge Distillation (AKD) and the Pseudo-Bag Memory Pool (PMP). AKD mitigates catastrophic forgetting by focusing on retaining attention layer knowledge between learning sessions, while PMP reduces the memory footprint by selectively storing only the most informative patches, or "pseudo-bags" from WSIs. Experimental evaluations demonstrate that our method significantly improves both accuracy and memory efficiency on diverse WSI datasets, outperforming current state-of-the-art CL methods. This work provides a foundation for CL in large-scale, weakly annotated clinical datasets, paving the way for more adaptable and resilient diagnostic models.

****

Beyond Image Classification: A Video Benchmark and Dual-Branch Hybrid Discrimination Framework for Compositional Zero-Shot Learning

Dongyao Jiang, Haodong Jing, Yongqiang Ma, Nanning Zheng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9860-9869

Human reasoning naturally combines concepts to identify unseen compositions, a capability that Compositional Zero-Shot Learning (CZSL) aims to replicate in machine learning models. However, we observe that focusing solely on typical image classification tasks in CZSL may limit models' compositional generalization poten

tial. To address this, we introduce C-EgoExo, a video-based benchmark, along wit h a compositional action recognition task to enable more comprehensive evaluatio ns. Inspired by human reasoning processes, we propose a Dual-branch Hybrid Discr imination (DHD) framework, featuring two branches that decode visual inputs in d istinct observation sequences. Through a cross-attention mechanism and a context ual dependency encoder, DHD effectively mitigates challenges posed by conditiona l variance. We further design a Copula-based orthogonal decoding loss to counter act contextual interference in primitive decoding. Our approach demonstrates out standing performance across diverse CZSL tasks, excelling in both image-based an d video-based modalities and in attribute-object and action-object compositions, setting a new benchmark for CZSL evaluation.

****

ABBSPO: Adaptive Bounding Box Scaling and Symmetric Prior based Orientation Pred iction for Detecting Aerial Image Objects

Woojin Lee, Hyugjae Chang, Jaeho Moon, Jaehyup Lee, Munchurl Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8848-8 858

Weakly supervised Oriented Object Detection (WS-OOD) has gained attention as a c ost-effective alternative to fully supervised methods, providing efficiency and high accuracy. Among weakly supervised approaches, horizontal bounding box (HBox) supervised OOD stands out for its ability to directly leverage existing HBox a nnotations while achieving the highest accuracy under weak supervision settings. This paper introduces adaptive bounding box scaling and symmetry-prior-based or ientation prediction, called ABBSPO that is a framework for WS-OOD. Our ABBSPO a ddresses the limitations of previous HBox-supervised OOD methods, which compare ground truth (GT) HBoxes directly with predicted RBoxes' minimum circumscribed r ectangles, often leading to inaccuracies. To overcome this, we propose: (i) Adap tive Bounding Box Scaling (ABBS) that appropriately scales the GT HBoxes to opti mize for the size of each predicted RBox, ensuring more accurate prediction for RBoxes' scales; and (ii) a Symmetric Prior Angle (SPA) loss that uses the inhere nt symmetry of aerial objects for self-supervised learning, addressing the issue in previous methods where learning fails if they consistently make incorrect pr edictions for all three augmented views (original, rotated, and flipped). Extens ive experimental results demonstrate that our ABBSPO achieves state-of-the-art r esults, outperforming existing methods.

****

Decoupled Distillation to Erase: A General Unlearning Method for Any Class-centr ic Tasks

Yu Zhou, Dian Zheng, Qijie Mo, Renjie Lu, Kun-Yu Lin, Wei-Shi Zheng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 203 50-20359

In this work, we present DEcoupLEd Distillation To Erase (DELETE), a general and strong unlearning method for any class-centric tasks. To derive this, we first propose a theoretical framework to analyze the general form of unlearning loss a nd decompose it into forgetting and retention terms. Through the theoretical fra mework, we point out that a class of previous methods could be mainly formulated as a loss that implicitly optimizes the forgetting term while lacking supervisi on for the retention term, disturbing the distribution of pre-trained model and struggling to adequately preserve knowledge of the remaining classes. To address it, we refine the retention term using "dark knowledge" and propose a mask dist illation unlearning method. By applying a mask to separate forgetting logits fro m retention logits, our approach optimizes both the forgetting and refined reten tion components simultaneously, retaining knowledge of the remaining classes whi le ensuring thorough forgetting of the target class. Without access to the remai ning data or intervention (i.e., used in some works), we achieve state-of-the-ar t performance across various benchmarks. What's more, DELETE is a general soluti on that can be applied to various downstream tasks, including face recognition, backdoor defense, and semantic segmentation with great performance.

****

TAET: Two-Stage Adversarial Equalization Training on Long-Tailed Distributions

Wang Yu-Hang, Junkang Guo, Aolei Liu, Kaihao Wang, Zaitong Wu, Zhenyu Liu, Wenfe i Yin, Jian Liu; Proceedings of the Computer Vision and Pattern Recognition Conf erence (CVPR), 2025, pp. 15476-15485
Adversarial robustness remains a significant challenge in deploying deep neural networks for real-world applications. While adversarial training is widely ackno wledged as a promising defense strategy, most existing studies primarily focus o n balanced datasets, neglecting the fact that real-world data often exhibit a lo ng-tailed distribution, which introduces substantial challenges to robustness. I n this paper, we provide an in-depth analysis of adversarial training in the con text of long-tailed distributions and identify the limitations of the current st ate-of-the-art method, AT-BSL, in achieving robust performance under such condit ions. To address these challenges, we propose a novel training framework, TAET, which incorporates an initial stabilization phase followed by a stratified, equa lization adversarial training phase. Furthermore, prior work on long-tailed robu stness has largely overlooked a crucial evaluation metric--Balanced Accuracy. To fill this gap, we introduce the concept of Balanced Robustness, a comprehensive metric that measures robustness specifically under long-tailed distributions. E xtensive experiments demonstrate that our method outperforms existing advanced d efenses, yielding significant improvements in both memory and computational effi ciency. We believe this work represents a substantial step forward in tackling r obustness challenges in real-world applications. Our paper code can be found at https://github.com/BuhuiOK/TAET-Two-Stage-Adversarial-Equalization-Training-on-L ong-Tailed-Distributions
****

Few-shot Personalized Scanpath Prediction
Ruoyu Xue, Jingyi Xu, Sounak Mondal, Hieu Le, Greg Zelinsky, Minh Hoai, Dimitris Samaras; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13497-13507
A personalized model for scanpath prediction provides insights into the visual p references and attention patterns of individual subjects. However, existing meth ods for training scanpath prediction models are data-intensive and cannot be eff ectively personalized to new individuals with only a few available examples. In this paper, we propose few-shot personalized scanpath prediction task (FS-PSP) a nd a novel method to address it, which aims to predict scanpaths for an unseen s ubject using minimal support data of that subject's scanpath behavior. The key t o our method's adaptability is the Subject-Embedding Network (SE-Net), specifica lly designed to capture unique, individualized representations for each subject' s scanpaths. SE-Net generates subject embeddings that effectively distinguish be tween subjects while minimizing variability among scanpaths from the same indivi dual. The personalized scanpath prediction model is then conditioned on these su bject embeddings to produce accurate, personalized results. Experiments on multi ple eye-tracking datasets demonstrate that our method excels in FS-PSP settings and does not require any fine-tuning steps at test time. Code is available at: h ttps://github.com/cvlab-stonybrook/few-shot-scanpath
****

Do Your Best and Get Enough Rest for Continual Learning
Hankyul Kang, Gregor Seifer, Donghyun Lee, Jongbin Ryu; Proceedings of the Compu ter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10077-10086
According to the forgetting curve theory, we can enhance memory retention by lea rning extensive data and taking adequate rest. This means that in order to effec tively retain new knowledge, it is essential to learn it thoroughly and ensure s ufficient rest so that our brain can memorize without forgetting. The main takea way from this theory is that learning extensive data at once necessitates suffic ient rest before learning the same data again. This aspect of human long-term me mory retention can be effectively utilized to address the continual learning of neural networks. Retaining new knowledge for a long period of time without catas trophic forgetting is the critical problem of continual learning. Therefore, bas ed on Ebbinghaus' theory, we introduce the view-batch model that adjusts the lea rning schedules to optimize the recall interval between retraining the same samp les. The proposed view-batch model allows the network to get enough rest to lear

n extensive knowledge from the same samples with a recall interval of sufficient length. To this end, we specifically present two approaches: 1) a replay method that guarantees the optimal recall interval, and 2) a self-supervised learning that acquires extensive knowledge from a single training sample at a time. We empirically show that these approaches of our method are aligned with the forgetting curve theory, which can enhance long-term memory. In our experiments, we also demonstrate that our method significantly improves many state-of-the-art continual learning methods in various protocols and scenarios. We open-source this project at https://github.com/hankyul2/ViewBatchModel.
****

Enhancing Few-Shot Class-Incremental Learning via Training-Free Bi-Level Modality Calibration
Yiyang Chen, Tianyu Ding, Lei Wang, Jing Huo, Yang Gao, Wenbin Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9881-9890
Few-shot Class-Incremental Learning (FSCIL) challenges models to adapt to new classes with limited samples, presenting greater difficulties than traditional classincremental learning. While existing approaches rely heavily on visual models and require additional training during base or incremental phases, we propose a training-free framework that leverages pre-trained visual-language models like CLIP. At the core of our approach is a novel Bilevel Modality Calibration (BiMC) strategy. Our framework initially performs intra-modal calibration, combining LLM-generated fine-grained category descriptions with visual prototypes from the base session to achieve precise classifier estimation. This is further complemented by inter-modal calibration that fuses pre-trained linguistic knowledge with task-specific visual priors to mitigate modality-specific biases. To enhance prediction robustness, we introduce additional metrics and strategies that maximize the utilization of limited data. Extensive experimental results demonstrate that our approach significantly outperforms existing methods. Code is available at: https://github.com/yychen016/BiMC.
****

MUSt3R: Multi-view Network for Stereo 3D Reconstruction
Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, Vincent Leroy; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1050-1060
DUSt3R introduced a novel paradigm in geometric computer vision by proposing a model that can provide dense and unconstrained Stereo 3D Reconstruction of arbitrary image collections with no prior information about camera calibration nor viewpoint poses. Under the hood, however, DUSt3R processes image pairs, regressing local 3D reconstructions that need to be aligned in a global coordinate system. The number of pairs, growing quadratically, is an inherent limitation that becomes especially concerning for robust and fast optimization in the case of large image collections. In this paper, we propose an extension of DUSt3R from pairs to multiple views, that addresses all aforementioned concerns. Indeed, we propose a Multi-view Network for Stereo 3D Reconstruction, or MUSt3R, that modifies the DUSt3R architecture by making it symmetric and extending it to directly predict 3D structure for all views in a common coordinate frame. Second, we entail the model with a multi-layer memory mechanism which allows to reduce the computational complexity and to scale the reconstruction to large collections, inferring thousands of 3D pointmaps at high frame-rates with limited added complexity. The framework is designed to perform 3D reconstruction both offline and online, and hence can be seamlessly applied to SfM and visual SLAM scenarios showing state-of-the-art performance on various 3D downstream tasks, including uncalibrated Visual Odometry, relative camera pose, scale and focal estimation, 3D reconstruction and multi-view depth estimation.
****

Hybrid-Level Instruction Injection for Video Token Compression in Multi-modal Large Language Models
Zhihang Liu, Chen-Wei Xie, Pandeng Li, Liming Zhao, Longxiang Tang, Yun Zheng, Chuanbin Liu, Hongtao Xie; Proceedings of the Computer Vision and Pattern Recogni

tion Conference (CVPR), 2025, pp. 8568-8578
Recent Multi-modal Large Language Models (MLLMs) have been challenged by the computational overhead resulting from massive video frames, often alleviated through compression strategies. However, the visual content is not equally contributed to user instructions, existing strategies (e.g., average pool) inevitably lead to the loss of potentially useful information. To tackle this, we propose the Hybrid-level Instruction Injection Strategy for Conditional Token Compression in MLLMs (HICom), utilizing the instruction as a condition to guide the compression from both local and global levels. This encourages the compression to retain the maximum amount of user-focused information while reducing visual tokens to minimize computational burden. Specifically, the instruction condition is injected into the grouped visual tokens at the local level and the learnable tokens at the global level, and we conduct the attention mechanism to complete the conditional compression. From the hybrid-level compression, the instruction-relevant visual parts are highlighted while the temporal-spatial structure is also preserved for easier understanding of LLMs. To further unleash the potential of HICom, we introduce a new conditional pre-training stage with our proposed dataset HICom-248K. Experiments show that our HICom can obtain distinguished video understanding ability with fewer tokens, increasing the performance by 2.43% average on three multiple-choice QA benchmarks and saving 78.8% tokens compared with the SOTA method. The code is available at https://github.com/lntzm/HICom.
****

Mamba4D: Efficient 4D Point Cloud Video Understanding with Disentangled Spatial-Temporal State Space Models
Jiuming Liu, Jinru Han, Lihao Liu, Angelica I. Aviles-Rivero, Chaokang Jiang, Zhe Liu, Hesheng Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17626-17636
Point cloud videos can faithfully capture real-world spatial geometries and temporal dynamics, which are essential for enabling intelligent agents to understand the dynamically changing world. However, designing an effective 4D backbone remains challenging, mainly due to the irregular and unordered distribution of points and temporal inconsistencies across frames. Also, recent transformer-based 4D backbones commonly suffer from large computational costs due to their quadratic complexity, particularly for long video sequences. To address these challenges, we propose a novel point cloud video understanding backbone purely based on the State Space Models (SSMs). Specifically, we first disentangle space and time in 4D video sequences and then establish the spatio-temporal correlation with the unified spatial-temporal Mamba blocks. The Intra-frame Spatial Mamba module is developed to encode locally similar geometric structures within a certain temporal stride. Subsequently, locally correlated tokens are delivered to the Inter-frame Temporal Mamba module, which integrates long-term point features across the entire video with linear complexity. Our proposed Mamba4D achieves competitive performance on the MSR-Action3D action recognition (+10.4% accuracy), HOI4D action segmentation (+0.7 F1 Score), and Synthia4D semantic segmentation (+0.19 mIoU) datasets. Mamba4D also has a significant efficiency improvement, especially for long video sequences, with 87.5% GPU memory reduction and 5.36 times speed-up. Codes are released at https://github.com/IRMVLab/Mamba4D.
****

Mamba as a Bridge: Where Vision Foundation Models Meet Vision Language Models for Domain-Generalized Semantic Segmentation
Xin Zhang, Robby T. Tan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14527-14537
Vision Foundation Models (VFMs) and Vision-Language Models (VLMs) have gained traction in Domain Generalized Semantic Segmentation (DGSS) due to their strong generalization capabilities. However, existing DGSS methods often rely exclusively on either VFMs or VLMs, overlooking their complementary strengths. VFMs (e.g., DINOv2) excel at capturing fine-grained features, while VLMs (e.g., CLIP) provide robust text alignment but struggle with coarse granularity. Despite their complementary strengths, effectively integrating VFMs and VLMs with attention mechanisms is challenging, as the increased patch tokens complicate long-sequence mode

ling. To address this, we propose MFuser, a novel Mamba-based fusion framework t
hat efficiently combines the strengths of VFMs and VLMs while maintaining linear
 scalability in sequence length. MFuser consists of two key components: MVFuser,
 which acts as a co-adapter to jointly fine-tune the two models by capturing bot
h sequential and spatial dynamics; and MTEnhancer, a hybrid attention-Mamba modu
le that refines text embeddings by incorporating image priors. Our approach achi
eves precise feature locality and strong text alignment without incurring signif
icant computational overhead. Extensive experiments demonstrate that MFuser sign
ificantly outperforms state-of-the-art DGSS methods, achieving 68.20 mIoU on syn
thetic-to-real and 71.87 mIoU on real-to-real benchmarks. The code is available
at https://github.com/devinxzhang/MFuser.
****

Vision-Guided Action: Enhancing 3D Human Motion Prediction with Gaze-informed Af
fordance in 3D Scenes

Ting Yu, Yi Lin, Jun Yu, Zhenyu Lou, Qiongjie Cui; Proceedings of the Computer V
ision and Pattern Recognition Conference (CVPR), 2025, pp. 12335-12346

Recent advances in human motion prediction (HMP) have shifted focus from isolate
d motion data to integrating human-scene correlations. In particular, the latest
 methods leverage human gaze points, using their spatial coordinates to indicate
 intent--where a person might move within a 3D environment. Despite promising tr
ajectory results, these methods often produce inaccurate poses by overlooking th
e semantic implications of gaze, specifically the affordances of observed object
s, which indicate the possible interactions. To address this, we propose GAP3DS,
 an affordance-aware HMP model that utilizes gaze-informed object affordances to
 improve HMP in complex 3D environments. GAP3DS incorporates a gaze-guided affor
dance learner to identify relevant objects in the scene and infer their affordan
ces based on human gaze, thus contextualizing future human-object interactions.
This affordance information, enriched with visual features and gaze data, condit
ions the generation of multiple human-object interaction poses, which are subseq
uently decoded into final motion predictions. Extensive experiments on two real-
world datasets demonstrate that GAP3DS outperforms state-of-the-art methods in b
oth trajectory and pose accuracy, producing more physically consistent and conte
xtually grounded predictions. For more details and code, please refer to the pro
ject page.
****

LOGICZSL: Exploring Logic-induced Representation for Compositional Zero-shot Lea
rning

Peng Wu, Xiankai Lu, Hao Hu, Yongqin Xian, Jianbing Shen, Wenguan Wang; Proceedi
ngs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp.
30301-30311

Compositional zero-shot learning (CZSL) aims to recognize unseen attribute-objec
t compositions by learning the primitive concepts (*i.e.*, attribute and object)
 from the training set. While recent works achieve impressive results in CZSL by
 leveraging large vision-language models like CLIP, they ignore the rich semanti
c relationships between primitive concepts and their compositions. In this work,
 we propose LOGICZSL, a novel logic-induced learning framework to explicitly mod
el the semantic relationships. Our logic-induced learning framework formulates t
he relational knowledge constructed from large language models as a set of logic
 rules, and grounds them onto the training data. Our logic-induced losses are co
mplementary to the widely used CZSL losses, therefore can be employed to inject
the semantic information into any existing CZSL methods. Extensive experimental
results show that our method brings significant performance improvements across
diverse datasets (*i.e.*, CGQA, UT-Zappos50K, MIT-States) with strong CLIP-based
 methods and settings (*i.e.*, Close World, Open World).
****

ChainHOI: Joint-based Kinematic Chain Modeling for Human-Object Interaction Gene
ration

Ling-An Zeng, Guohong Huang, Yi-Lin Wei, Shengbo Gu, Yu-Ming Tang, Jingke Meng,
Wei-Shi Zheng; Proceedings of the Computer Vision and Pattern Recognition Confer
ence (CVPR), 2025, pp. 12358-12369

We propose ChainHOI, a novel approach for text-driven human-object interaction (HOI) generation that explicitly models interactions at both the joint and kineti c chain levels. Unlike existing methods that implicitly model interactions us ing full-body poses as tokens, we argue that explicitly modeling joint-level int eractions is more natural and effective for generating realistic HOIs, as it dir ectly captures the geometric and semantic relationships between joints, rather t han modeling interactions in the latent pose space. To this end, ChainHOI int roduces a novel joint graph to capture potential interactions with objects, and a Generative Spatiotemporal Graph Convolution Network to explicitly model intera ctions at the joint level. Furthermore, we propose a Kinematics-based Intera ction Module that explicitly models interactions at the kinetic chain level, ens uring more realistic and biomechanically coherent motions. Evaluations on two public datasets demonstrate that ChainHOI significantly outperforms previous me thods, generating more realistic, and semantically consistent HOIs.

****

CLOC: Contrastive Learning for Ordinal Classification with Multi-Margin N-pair L oss

Dileepa Pitawela, Gustavo Carneiro, Hsiang-Ting Chen; Proceedings of the Compute r Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15538-15548

In ordinal classification, misclassifying neighboring ranks is common, yet the c onsequences of these errors are not the same.For example, misclassifying benign tumor categories is less consequential, compared to an error at the pre-cancerou s to cancerous threshold, which could profoundly influence treatment choices. De spite this, existing ordinal classification methods do not account for the varyi ng importance of these margins, treating all neighboring classes as equally sign ificant. To address this limitation, we propose CLOC, a new margin-based contras tive learning method for ordinal classification that learns an ordered represent ation based on the optimization of multiple margins with a novel multi-margin n- pair loss (MMNP).CLOC enables flexible decision boundaries across key adjacent c ategories, facilitating smooth transitions between classes and reducing the risk of overfitting to biases present in the training data.We provide empirical disc ussion regarding the properties of MMNP and show experimental results on five re al-world image datasets (Adience, Historical Colour Image Dating, Knee Osteoarth ritis, Indian Diabetic Retinopathy Image, and Breast Carcinoma Subtyping) and on e synthetic dataset simulating clinical decision bias.Our results demonstrate th at CLOC outperforms existing ordinal classification methods and show the interpr etability and controllability of CLOC in learning meaningful, ordered representa tions that align with clinical and practical needs.

****

Universal Actions for Enhanced Embodied Foundation Models

Jinliang Zheng, Jianxiong Li, Dongxiu Liu, Yinan Zheng, Zhihao Wang, Zhonghong O u, Yu Liu, Jingjing Liu, Ya-Qin Zhang, Xianyuan Zhan; Proceedings of the Compute r Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22508-22519

Training on diverse, internet-scale data is a key factor in the success of recen t large foundation models. Yet, using the same recipe for building embodied agen ts has faced noticeable difficulties. Despite the availability of many crowd-sou rced embodied datasets, their action spaces often exhibit significant heterogene ity due to distinct physical embodiment and control interfaces for different rob ots, causing substantial challenges in developing embodied foundation models usi ng cross-embodiment data. In this paper, we introduce UniAct, a new embodied fou ndation modeling framework operating in the Universal Action Space. Our learned universal actions capture the generic behaviors across diverse robots by exploit ing their shared structural features, and enable enhanced cross-domain data util ization and cross-embodiment generalizations by eliminating the notorious hetero geneity. Moreover, the universal actions can be efficiently translated back to h eterogeneous actionable commands by simply adding embodiment-specific details, f rom which fast adaptation to new robots becomes simple and straightforward. Our 0.5B instantiation of UniAct outperforms 14X larger SOTA embodied foundation mod els in extensive evaluations on various real-world and simulation robots, showca sing exceptional cross-embodiment control and adaptation capability, highlightin

g the crucial benefit of adopting universal actions.
****

ObjectMover: Generative Object Movement with Video Prior
Xin Yu, Tianyu Wang, Soo Ye Kim, Paul Guerrero, Xi Chen, Qing Liu, Zhe Lin, Xiaojuan Qi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17682-17691
Simple as it seems, moving an object to another location within an image is, in fact, a challenging image-editing task that requires re-harmonizing the lighting, adjusting the pose based on perspective, accurately filling occluded regions, and ensuring coherent synchronization of shadows and reflections while maintaining the object identity. In this paper, we present ObjectMover, a generative model that can perform object movement in highly challenging scenes. Our key insight is that we model this task as a sequence-to-sequence problem and fine-tune a video generation model to leverage its knowledge of consistent object generation across video frames. We show that with this approach, our model is able to adjust to complex real-world scenarios, handling extreme lighting harmonization and object effect movement. As large-scale data for object movement are unavailable, we construct a data generation pipeline using a modern game engine to synthesize high-quality data pairs. We further propose a multi-task learning strategy that enables training on real-world video data to improve the model generalization. Through extensive experiments, we demonstrate that ObjectMover achieves outstanding results and adapts well to real-world scenarios.
****

FaithDiff: Unleashing Diffusion Priors for Faithful Image Super-resolution
Junyang Chen, Jinshan Pan, Jiangxin Dong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28188-28197
Faithful image super-resolution (SR) not only needs to recover images that appear realistic, similar to image generation tasks, but also requires that the restored images maintain fidelity and structural consistency with the input. To this end, we propose a simple and effective method, named FaithDiff, to fully harness the impressive power of latent diffusion models (LDMs) for faithful image SR. In contrast to existing diffusion-based SR methods that freeze the diffusion model pre-trained on high-quality images, we propose to unleash the diffusion prior to identify useful information and recover faithful structures. As there exists a significant gap between the features of degraded inputs and the noisy latent from the diffusion model, we then develop an effective alignment module to explore useful features from degraded inputs to align well with the diffusion process. Considering the indispensable roles and interplay of the encoder and diffusion model in LDMs, we jointly fine-tune them in a unified optimization framework, facilitating the encoder to extract useful features that coincide with diffusion process. Extensive experimental results demonstrate that FaithDiff outperforms state-of-the-art methods, providing high-quality and faithful SR results.
****

MLLM-as-a-Judge for Image Safety without Human Labeling
Zhenting Wang, Shuming Hu, Shiyu Zhao, Xiaowen Lin, Felix Juefei-Xu, Zhuowei Li, Ligong Han, Harihar Subramanyam, Li Chen, Jianfa Chen, Nan Jiang, Lingjuan Lyu, Shiqing Ma, Dimitris N. Metaxas, Ankit Jain; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14657-14666
Image content safety has become a significant challenge with the rise of visual media on online platforms. Meanwhile, in the age of AI-generated content (AIGC), many image generation models are capable of producing harmful content, such as images containing sexual or violent material. Thus, it becomes crucial to identify such unsafe images based on established safety rules. Pre-trained Multimodal Large Language Models (MLLMs) offer potential in this regard, given their strong pattern recognition abilities. Existing approaches typically fine-tune MLLMs with human-labeled datasets, which however brings a series of drawbacks. First, relying on human annotators to label data following intricate and detailed guidelines is both expensive and labor-intensive. Furthermore, users of safety judgment systems may need to frequently update safety rules, making fine-tuning on human-based annotation more challenging. This raises the research question: Can we de

tect unsafe images by querying MLLMs in a zero-shot setting using a predefined s afety constitution (a set of safety rules)? Our research showed that simply quer ying pre-trained MLLMs does not yield satisfactory results. This lack of effecti veness stems from factors such as the subjectivity of safety rules, the complexi ty of lengthy constitutions, and the inherent biases in the models. To address t hese challenges, we propose a MLLM-based method includes objectifying safety rul es, assessing the relevance between rules and images, making quick judgments bas ed on debiased token probabilities with logically complete yet simplified precon dition chains for safety rules, and conducting more in-depth reasoning with casc aded chain-of-thought processes if necessary. Experiment results demonstrate tha t our method is highly effective for zero-shot image safety judgment tasks.
****

A New Statistical Model of Star Speckles for Learning to Detect and Characterize Exoplanets in Direct Imaging Observations
Théo Bodrito, Olivier Flasseur, Julien Mairal, Jean Ponce, Maud Langlois, Anne-M arie Lagrange; Proceedings of the Computer Vision and Pattern Recognition Confer ence (CVPR), 2025, pp. 1230-1240
The search for exoplanets is an active field in astronomy, with direct imaging a s one of the most challenging methods due to faint exoplanet signals buried with in stronger residual starlight. Successful detection requires advanced image pr ocessing to separate the exoplanet signal from this nuisance component. This pap er presents a novel statistical model that captures nuisance fluctuations using a multi-scale approach, leveraging problem symmetries and a joint spectral chann el representation grounded in physical principles. Our model integrates into an interpretable, end-to-end learnable framework for simultaneous exoplanet detect ion and flux estimation. The proposed algorithm is evaluated against the state of the art using datasets from the SPHERE instrument operating at the Very Large Telescope (VLT). It significantly improves the precision-recall trade-off, not ably on challenging datasets that are otherwise unusable by astronomers. The pr oposed approach is computationally efficient, robust to varying data quality, an d well suited for large-scale observational surveys.
****

Scene-agnostic Pose Regression for Visual Localization
Junwei Zheng, Ruiping Liu, Yufan Chen, Zhenfang Chen, Kailun Yang, Jiaming Zhang , Rainer Stiefelhagen; Proceedings of the Computer Vision and Pattern Recognitio n Conference (CVPR), 2025, pp. 27092-27102
Absolute Pose Regression (APR) predicts 6D camera poses but lacks the adaptabili ty to unknown environments without retraining, while Relative Pose Regression (R PR) generalizes better yet requires a large image retrieval database. Visual Odo metry (VO) generalizes well in unseen environments but suffers from accumulated error in open trajectories. To address this dilemma, we introduce a new task, Sc ene-agnostic Pose Regression (SPR), which can achieve accurate pose regression i n a flexible way while eliminating the need for retraining or databases. To benc hmark SPR, we created a large-scale dataset, 360SPR, with over 200K photorealist ic panoramas, 3.6M pinhole images and camera poses in 270 scenes at three differ ent sensor heights. Furthermore, a SPR-Mamba model is initially proposed to addr ess SPR in a dual-branch manner. Extensive experiments and studies demonstrate t he effectiveness of our SPR paradigm, dataset, and model. In the unknown scenes of both 360SPR and 360Loc datasets, our method consistently outperforms APR, RPR and VO. The dataset and code are available at SPR.
****

Learning to Filter Outlier Edges in Global SfM
Nicole Damblon, Marc Pollefeys, Daniel Barath; Proceedings of the Computer Visio n and Pattern Recognition Conference (CVPR), 2025, pp. 11558-11568
We present a novel approach to enhance camera pose estimation in global Structur e-from-Motion (SfM) frameworks by filtering inaccurate pose graph edges - repres enting relative translation estimates - before applying translation averaging. I n SfM, pose graph vertices represent images, and edges represent relative poses (rotations and translations) between cameras. We reformulate the edge filtering problem as a vertex filtering in the dual graph, specifically, a line graph wher

e vertices correspond to edges in the original graph and edges correspond to cam
eras. Utilizing this representation, we frame the problem as a binary classifica
tion over nodes in the dual graph. To identify outlier edges, we employ a Transf
ormer-based architecture. To overcome the challenge of memory overflow caused by
 converting to a line graph, we introduce a clustering-based graph processing ap
proach, enabling our method to be applied to arbitrarily large pose graphs. Our
method outperforms existing relative translation filtering techniques in terms o
f camera position accuracy and can be seamlessly integrated with other filters.
The code is available at https://github.com/DmblnNicole/LFOE-GlobalSfM.
****

Divide and Conquer: Heterogeneous Noise Integration for Diffusion-based Adversar
ial Purification
Gaozheng Pei, Shaojie Lyu, Gong Chen, Ke Ma, Qianqian Xu, Yingfei Sun, Qingming
Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (CV
PR), 2025, pp. 29268-29277
Existing diffusion-based purification methods aim to disrupt adversarial perturb
ations by introducing a certain amount of noise through a forward diffusion proc
ess, followed by a reverse process to recover clean examples. However, this appr
oach is fundamentally flawed: the uniform operation of the forward process acros
s all pixels compromises normal pixels while attempting to combat adversarial pe
rturbations, resulting in the target model producing incorrect predictions. Simp
ly relying on low-intensity noise is insufficient for effective defense. To addr
ess this critical issue, we implement a heterogeneous purification strategy grou
nded in the interpretability of neural networks. Our method decisively applies h
igher-intensity noise to specific pixels that the target model focuses on while
the remaining pixels are subjected to only low-intensity noise. This requirement
 motivates us to redesign the sampling process of the diffusion model, allowing
for the effective removal of varying noise levels. Furthermore, to evaluate our
method against strong adaptative attack, our proposed method sharply reduces tim
e cost and memory usage through a single-step resampling. The empirical evidence
 from extensive experiments across three datasets demonstrates that our method o
utperforms most current adversarial training and purification techniques by a su
bstantial margin.
****

SEC-Prompt:SEmantic Complementary Prompting for Few-Shot Class-Incremental Learn
ing
Ye Liu, Meng Yang; Proceedings of the Computer Vision and Pattern Recognition Co
nference (CVPR), 2025, pp. 25643-25656
Few-shot class-incremental learning (FSCIL) presents a significant challenge in
machine learning, requiring models to integrate new classes from limited example
s while preserving performance on previously learned classes. Recently, prompt-b
ased CIL approaches leverage ample data to train prompts, effectively mitigating
 catastrophic forgetting. However, these methods do not account for the semantic
 features embedded in prompts, exacerbating the plasticity-stability dilemma in
few-shot incremental learning. In this paper, we propose a novel and simple fram
ework named SEmantic Complementary Prompt(SEC-Prompt), which learns two sets of
semantically complementary prompts based on an adaptive query: discriminative pr
ompts(D-Prompt) and non-discriminative prompts(ND-Prompt). D-Prompt enhances the
 separation of class-specific feature distributions by strengthening key discrim
inative features, while ND-Prompt balances non-discriminative information to pro
mote generalization to novel classes. To efficiently learn high-quality knowledg
e from limited samples, we leverage ND-Prompt for data augmentation to increase
sample diversity and introduce Prompt Clustering Loss to prevent noise contamina
tion in D-Prompt, ensuring robust discriminative feature learning and improved g
eneralization. Our experimental results showcase state-of-the-art performance ac
ross three benchmark datasets, including CIFAR100, ImageNet-R and CUB datasets.
****

LiMoE: Mixture of LiDAR Representation Learners from Automotive Scenes
Xiang Xu, Lingdong Kong, Hui Shuai, Liang Pan, Ziwei Liu, Qingshan Liu; Proceedi
ngs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp.

27368-27379
LiDAR data pretraining offers a promising approach to leveraging large-scale, readily available datasets for enhanced data utilization. However, existing methods predominantly focus on sparse voxel representation, overlooking the complementary attributes provided by other LiDAR representations. In this work, we propose LiMoE, a framework that integrates the Mixture of Experts (MoE) paradigm into LiDAR data representation learning to synergistically combine multiple representations, such as range images, sparse voxels, and raw points. Our approach consists of three stages: i) Image-to-LiDAR Pretraining, which transfers prior knowledge from images to point clouds across different representations; ii) Contrastive Mixture Learning (CML), which uses MoE to adaptively activate relevant attributes from each representation and distills these mixed features into a unified 3D network; iii) Semantic Mixture Supervision (SMS), which combines semantic logits from multiple representations to boost downstream segmentation performance. Extensive experiments across eleven large-scale LiDAR datasets demonstrate our effectiveness and superiority. The code has been made publicly accessible.
****

CoT-VLA: Visual Chain-of-Thought Reasoning for Vision-Language-Action Models
Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, Ankur Handa, Tsung-Yi Lin, Gordon Wetzstein, Ming-Yu Liu, Donglai Xiang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1702-1713
Vision-language-action models (VLAs) have shown potential in leveraging pretrained vision-language models and diverse robot demonstrations for learning generalizable sensorimotor control. While this paradigm effectively utilizes large-scale data from both robotic and non-robotic sources, current VLAs primarily focus on direct input--output mappings, lacking the intermediate reasoning steps crucial for complex manipulation tasks. As a result, existing VLAs lack temporal planning or reasoning capabilities. In this paper, we introduce a method that incorporates explicit visual chain-of-thought (CoT) reasoning into vision-language-action models (VLAs) by predicting future image frames autoregressively as visual goals before generating a short action sequence to achieve these goals. We introduce CoT-VLA, a state-of-the-art 7B VLA that can understand and generate visual and action tokens. Our experimental results demonstrate that CoT-VLA achieves strong performance, outperforming the state-of-the-art VLA model by 17% in real-world manipulation tasks and 6% in simulation benchmarks. Videos are available at: https://cot-vla.github.io/.
****

WAVE: Weight Templates for Adaptive Initialization of Variable-sized Models
Fu Feng, Yucheng Xie, Jing Wang, Xin Geng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4819-4828
The growing complexity of model parameters underscores the significance of pre-trained models. However, deployment constraints often necessitate models of varying sizes, exposing limitations in the conventional pre-training and fine-tuning paradigm, particularly when target model sizes are incompatible with pre-trained ones. To address this challenge, we propose WAVE, a novel approach that reformulates variable-sized model initialization from a multi-task perspective, where initializing each model size is treated as a distinct task. WAVE employs shared, size-agnostic weight templates alongside size-specific weight scalers to achieve consistent initialization across various model sizes. These weight templates, constructed within the Learngene framework, integrate knowledge from pre-trained models through a distillation process constrained by Kronecker-based rules. Target models are then initialized by concatenating and weighting these templates, with adaptive connection rules established by lightweight weight scalers, whose parameters are learned from minimal training data. Extensive experiments demonstrate the efficiency of WAVE, achieving state-of-the-art performance in initializing models of various depth and width. The knowledge encapsulated in weight templates is also task-agnostic, allowing for seamless transfer across diverse downstream datasets. Code will be made available at https://github.com/fu-feng/WAVE.
****

Forensics Adapter: Adapting CLIP for Generalizable Face Forgery Detection
Xinjie Cui, Yuezun Li, Ao Luo, Jiaran Zhou, Junyu Dong; Proceedings of the Compu ter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19207-19217
We describe the Forensics Adapter, an adapter network designed to transform CLIP into an effective and generalizable face forgery detector. Although CLIP is hig hly versatile, adapting it for face forgery detection is non-trivial as forgery-related knowledge is entangled with a wide range of unrelated knowledge. Existin g methods treat CLIP merely as a feature extractor, lacking task-specific adapta tion, which limits their effectiveness. To address this, we introduce an adapter to learn face forgery traces -- the blending boundaries unique to forged faces, guided by task-specific objectives. Then we enhance the CLIP visual tokens with a dedicated interaction strategy that communicates knowledge across CLIP and th e adapter. Since the adapter is alongside CLIP, its versatility is highly retain ed, naturally ensuring strong generalizability in face forgery detection. With o nly 5.7M trainable parameters, our method achieves a significant performance boo st, improving by approximately 7% on average across five standard datasets. We b elieve the proposed method can serve as a baseline for future CLIP-based face fo rgery detection methods.
****
KAC: Kolmogorov-Arnold Classifier for Continual Learning
Yusong Hu, Zichen Liang, Fei Yang, Qibin Hou, Xialei Liu, Ming-Ming Cheng; Proce edings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, p p. 15297-15307
Continual learning requires models to train continuously across consecutive task s without forgetting. Most existing methods utilize linear classifiers, which st ruggle to maintain a stable classification space while learning new tasks. Inspi red by the success of Kolmogorov-Arnold Networks (KAN) in preserving learning st ability during simple continual regression tasks, we set out to explore their po tential in more complex continual learning scenarios. In this paper, we introduc e the Kolmogorov-Arnold Classifier (KAC), a novel classifier developed for conti nual learning based on the KAN structure. We delve into the impact of KAN's spli ne functions and introduce Radial Basis Functions (RBF) for improved compatibili ty with continual learning. We replace linear classifiers with KAC in several re cent approaches and conduct experiments across various continual learning benchm arks, all of which demonstrate performance improvements, highlighting the effect iveness and robustness of KAC in continual learning.
****
PI-HMR: Towards Robust In-bed Temporal Human Shape Reconstruction with Contact P ressure Sensing
Ziyu Wu, Yufan Xiong, Mengting Niu, Fangting Xie, Quan Wan, Qijun Ying, Boyan Li u, Xiaohui Cai; Proceedings of the Computer Vision and Pattern Recognition Confe rence (CVPR), 2025, pp. 27739-27749
Long-term in-bed monitoring benefits automatic and real-time health management w ithin healthcare, and the advancement of human shape reconstruction technologies further enhances the representation and visualization of users' activity patter ns. However, existing technologies are primarily based on visual cues, facing se rious challenges in non-light-of-sight and privacy-sensitive in-bed scenes. Pres sure-sensing bedsheets offer a promising solution for real-time motion reconstru ction. Yet, limited exploration in model designs and data have hindered its furt her development. To tackle these issues, we propose a general framework that bri dges gaps in data annotation and model design. Firstly, we introduce SMPLify-IB, an optimization method that overcomes the depth ambiguity issue in top-view sce narios through gravity constraints, enabling generating high-quality 3D human sh ape annotations for in-bed datasets. Then we present PI-HMR, a temporal-based hu man shape estimator to regress meshes from pressure sequences. By integrating mu lti-scale feature fusion with high-pressure distribution and spatial position pr iors, PI-HMR outperforms SOTA methods with 17.01mm Mean-Per-Joint-Error decrease . This work provides a whole tool-chain to support the development of in-bed mon itoring with pressure contact sensing.
****

BOOTPLACE: Bootstrapped Object Placement with Detection Transformers
Hang Zhou, Xinxin Zuo, Rui Ma, Li Cheng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19294-19303
In this paper, we tackle the copy-paste image-to-image composition problem with a focus on object placement learning. Prior methods have leveraged generative models to reduce the reliance for dense supervision. However, this often limits their capacity to model complex data distributions. Alternatively, transformer networks with a sparse contrastive loss have been explored, but their over-relaxed regularization often leads to imprecise object placement. We introduce BOOTPLACE, a novel paradigm that formulates object placement as a placement-by-detection problem. Our approach begins by identifying suitable regions of interest for object placement. This is achieved by training a specialized detection transformer on object-subtracted backgrounds, enhanced with multi-object supervisions. It then semantically associates each target compositing object with detected regions based on their complementary characteristics. Through a boostrapped training approach applied to randomly object-subtracted images, our model enforces meaningful placements through extensive paired data augmentation. Experimental results on established benchmarks demonstrate BOOTPLACE's superior performance in object repositioning, markedly surpassing state-of-the-art baselines on Cityscapes and OPA datasets with notable improvements in IOU scores. Additional ablation studies further showcase the compositionality and generalizability of our approach, supported by user study evaluations. Code is available at https://github.com/RyanHangZhou/BOOTPLACE
****

CheckManual: A New Challenge and Benchmark for Manual-based Appliance Manipulation
Yuxing Long, Jiyao Zhang, Mingjie Pan, Tianshu Wu, Taewhan Kim, Hao Dong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22595-22604
Correct use of electrical appliances has significantly improved human life quality. Unlike simple tools that can be manipulated with common sense, different parts of electrical appliances have specific functions defined by manufacturers. If we want the robot to heat bread by microwave, we should enable them to review the microwave's manual first. From the manual, it can learn about component functions, interaction methods, and representative task steps about appliances. However, previous manual-related works remain limited to question-answering tasks while existing manipulation researchers ignore the manual's important role and fail to comprehend multi-page manuals. In this paper, we propose the first manual-based appliance manipulation benchmark CheckManual. Specifically, we design a large model-assisted human-revised data generation pipeline to create manuals based on CAD appliance models. With these manuals, we establish novel manual-based manipulation challenges, metrics, and simulator environments for model performance evaluation. Furthermore, we propose the first manual-based manipulation planning model ManualPlan to set up a group of baselines for the CheckManual benchmark.
****

CXPMRG-Bench: Pre-training and Benchmarking for X-ray Medical Report Generation on CheXpert Plus Dataset
Xiao Wang, Fuling Wang, Yuehang Li, Qingchuan Ma, Shiao Wang, Bo Jiang, Jin Tang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5123-5133
X-ray image-based medical report generation (MRG) is a pivotal area in artificial intelligence that can significantly reduce diagnostic burdens and patient wait times. Despite significant progress, we believe that the task has reached a bottleneck due to the limited benchmark datasets and the existing large models' insufficient capability enhancements in this specialized domain. Specifically, the recently released CheXpert Plus dataset lacks comparative evaluation algorithms and their results, providing only the dataset itself. This situation makes the training, evaluation, and comparison of subsequent algorithms challenging. Thus, we conduct a comprehensive benchmarking of existing mainstream X-ray report generation models and large language models (LLMs), on the CheXpert Plus dataset. We

believe that the proposed benchmark can provide a solid comparative basis for subsequent algorithms and serve as a guide for researchers to quickly grasp the state-of-the-art models in this field. More importantly, we propose a large model for the X-ray image report generation using a multi-stage pre-training strategy, including self-supervised autoregressive generation and Xray-report contrastive learning, and supervised fine-tuning. Extensive experimental results indicate that the autoregressive pre-training based on Mamba effectively encodes X-ray images, and the image-text contrastive pre-training further aligns the feature spaces, achieving better experimental results.

****

FASTer: Focal token Acquiring-and-Scaling Transformer for Long-term 3D Objection Detection

Chenxu Dang, ZaiPeng Duan, Pei An, Xinmin Zhang, Xuzhong Hu, Jie Ma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17029-17038

Recent top-performing temporal 3D detectors based on Lidars have increasingly adopted region-based paradigms. They first generate coarse proposals, followed by encoding and fusing regional features. However, indiscriminate sampling and fusion often overlook the varying contributions of individual points and lead to exponentially increased complexity as the number of input frames grows. Moreover, arbitrary result-level concatenation limits the global information extraction. In this paper, we propose a Focal Token Acquring-and-Scaling Transformer (FASTer), which dynamically selects focal tokens and condenses token sequences in an adaptive and lightweight manner. Emphasizing the contribution of individual tokens, we propose a simple but effective Adaptive Scaling mechanism to capture geometric contexts while sifting out focal points. Adaptively storing and processing only focal points in historical frames dramatically reduces the overall complexity. Furthermore, a novel Grouped Hierarchical Fusion strategy is proposed, progressively performing sequence scaling and Intra-Group Fusion operations to facilitate the exchange of global spatial and temporal information. Experiments on the Waymo Open Dataset demonstrate that our FASTer significantly outperforms other state-of-the-art detectors in both performance and efficiency while also exhibiting improved flexibility and robustness. The code is available at https://github.com/MSunDYY/FASTer.git.

****

SEEN-DA: SEmantic ENtropy guided Domain-aware Attention for Domain Adaptive Object Detection

Haochen Li, Rui Zhang, Hantao Yao, Xin Zhang, Yifan Hao, Xinkai Song, Shaohui Peng, Yongwei Zhao, Chen Zhao, Yanjun Wu, Ling Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25465-25475

Domain adaptive object detection (DAOD) aims to generalize detectors trained on an annotated source domain to an unlabelled target domain. Traditional works focus on aligning visual features between domains to extract domain-invariant knowledge, and recent VLM-based DAOD methods leverage semantic information provided by the textual encoder to supplement domain-specific features for each domain.However, they overlook the role of semantic information in guiding the learning of visual features that are beneficial for adaptation.To solve the problem, we propose semantic entropy to quantify the semantic information contained in visual features, and design SEmantic ENtropy guided Domain-aware Attention (SEEN-DA) to adaptively refine visual features with the semantic information of two domains.Semantic entropy reflects the importance of features based on semantic information, which can serve as attention to select discriminative visual features and suppress semantically irrelevant redundant information.Guided by semantic entropy, we introduce domain-aware attention modules into the visual encoder in SEEN-DA.It utilizes an inter-domain attention branch to extract domain-invariant features and eliminate redundant information, and an intra-domain attention branch to supplement the domain-specific semantic information discriminative on each domain.Comprehensive experiments validate the effectiveness of SEEN-DA, demonstrating significant improvements in cross-domain object detection performance.

****

Event-Equalized Dense Video Captioning
Kangyi Wu, Pengna Li, Jingwen Fu, Yizhe Li, Yang Wu, Yuhan Liu, Jinjun Wang, San ping Zhou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8417-8427
Dense video captioning aims to localize and caption all events in arbitrary untr immed videos. Although previous methods have achieved appealing results, they st ill face the issue of temporal bias, i.e, models tend to focus more on events wi th certain temporal characteristics. Specifically, 1) the temporal distribution of events in training datasets is uneven. Models trained on these datasets will pay less attention to out-of-distribution events. 2) long-duration events have m ore frame features than short ones and will attract more attention. To address t his, we argue that events, with varying temporal characteristics, should be trea ted equally when it comes to dense video captioning. Intuitively, different even ts tend to have distinct visual differences due to varied camera views, backgrou nds, or subjects. Inspired by that, we intend to utilize visual features to have an approximate perception of possible events and pay equal attention to them. I n this paper, we introduce a simple but effective framework, called Event-Equali zed Dense Video Captioning(E^2DVC) to overcome the temporal bias and treat all p ossible events equally. Specifically, an event perception module(EPM) is propose d to do uneven clustering on visual frame features to generate pseudo-events. We enforce the model's attention to these pseudo-events through the pseudo-event i nitialization module(PEI). A novel event-enhanced encoder(EEE) is also devised t o enhance the model's ability to explore frame-frame and frame-event relationshi ps. Experimental results validate the effectiveness of the proposed methods.
****

Geometry-guided Online 3D Video Synthesis with Multi-View Temporal Consistency
Hyunho Ha, Lei Xiao, Christian Richardt, Thu Nguyen-Phuoc, Changil Kim, Min H. K im, Douglas Lanman, Numair Khan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11275-11285
We introduce a novel geometry-guided online video view synthesis method with enh anced view and temporal consistency. Traditional approaches achieve high-quality synthesis from dense multi-view camera setups but require significant computati onal resources. In contrast, selective-input methods reduce this cost but often compromise quality, leading to multi-view and temporal inconsistencies such as f lickering artifacts. Our method addresses this challenge to deliver efficient, h igh-quality novel-view synthesis with view and temporal consistency. The key inn ovation of our approach lies in using global geometry to guide an image-based re ndering pipeline. To accomplish this, we progressively refine depth maps using c olor difference masks across time. These depth maps are then accumulated through truncated signed distance fields in the synthesized view's image space. This de pth representation is view and temporally consistent, and is used to guide a pre -trained blending network that fuses multiple forward-rendered input-view images . Thus, the network is encouraged to output geometrically consistent synthesis r esults across multiple views and time. Our approach achieves consistent, high-qu ality video synthesis, while running efficiently in an online manner.
****

EDCFlow: Exploring Temporally Dense Difference Maps for Event-based Optical Flow Estimation
Daikun Liu, Lei Cheng, Teng Wang, Changyin Sun; Proceedings of the Computer Visi on and Pattern Recognition Conference (CVPR), 2025, pp. 1984-1993
Recent learning-based methods for event-based optical flow estimation utilize co st volumes for pixel matching but suffer from redundant computations and limited scalability to higher resolutions for flow refinement. In this work, we take ad vantage of the complementarity between temporally dense feature differences of a djacent event frames and cost volume and present a lightweight event-based optic al flow network (EDCFlow) to achieve high-quality flow estimation at a higher re solution. Specifically, an attention-based multi-scale temporal feature differen ce layer is developed to capture diverse motion patterns at high resolution in a computation-efficient manner. An adaptive fusion of high-resolution difference motion features and low-resolution correlation motion features is performed to e

nhance motion representation and model generalization. Notably, EDCFlow can serv e as a plug-and-play refinement module for RAFT-like event-based methods to enha nce flow details. Extensive experiments demonstrate that EDCFlow achieves better performance with lower complexity compared to existing methods, offering superi or generalization. Codes and models will be available at here.
****

Point2RBox-v2: Rethinking Point-supervised Oriented Object Detection with Spatia l Layout Among Instances
Yi Yu, Botao Ren, Peiyuan Zhang, Mingxin Liu, Junwei Luo, Shaofeng Zhang, Feipen g Da, Junchi Yan, Xue Yang; Proceedings of the Computer Vision and Pattern Recog nition Conference (CVPR), 2025, pp. 19283-19293
With the rapidly increasing demand for oriented object detection (OOD), recent r esearch involving weakly-supervised detectors for learning OOD from point annota tions has gained great attention. In this paper, we rethink this challenging tas k setting with the layout among instances and present Point2RBox-v2. At the core are three principles: 1) Gaussian overlap loss. It learns an upper bound for ea ch instance by treating objects as 2D Gaussian distributions and minimizing thei r overlap. 2) Voronoi watershed loss. It learns a lower bound for each instance through watershed on Voronoi tessellation. 3) Consistency loss. It learns the si ze/rotation variation between two output sets with respect to an input image and its augmented view. Supplemented by a few devised techniques, e.g. edge loss an d copy-paste, the detector is further enhanced. To our best knowledge, Point2RBo x-v2 is the first approach to explore the spatial layout among instances for lea rning point-supervised OOD. Our solution is elegant and lightweight, yet it is e xpected to give a competitive performance especially in densely packed scenes: 6 2.61%/86.15%/34.71% on DOTA/HRSC/FAIR1M.
****

LibraGrad: Balancing Gradient Flow for Universally Better Vision Transformer Att ributions
Faridoun Mehri, Mahdieh Soleymani Baghshah, Mohammad Taher Pilehvar; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 67-78
Why do gradient-based explanations struggle with Transformers, and how can we im prove them? We identify gradient flow imbalances in Transformers that violate Fu llGrad-completeness, a critical property for attribution faithfulness that CNNs naturally possess. To address this issue, we introduce LibraGrad--a theoreticall y grounded post-hoc approach that corrects gradient imbalances through pruning a nd scaling of backward paths, without changing the forward pass or adding comput ational overhead. We evaluate LibraGrad using three metric families: Faithfulnes s, which quantifies prediction changes under perturbations of the most and least relevant features; Completeness Error, which measures attribution conservation relative to model outputs; and Segmentation AP, which assesses alignment with hu man perception. Extensive experiments across 8 architectures, 4 model sizes, and 5 datasets show that LibraGrad universally enhances gradient-based methods, out performing existing white-box methods--including Transformer-specific approaches --across all metrics. We demonstrate superior qualitative results through two co mplementary evaluations: precise text-prompted region highlighting on CLIP model s and accurate class discrimination between co-occurring animals on ImageNet-fin etuned models--two settings on which existing methods often struggle. LibraGrad is effective even on the attention-free MLP-Mixer architecture, indicating poten tial for extension to other modern architectures. Our code is freely available a t https://nightmachinery.github.io/LibraGrad/.
****

Blind Bitstream-corrupted Video Recovery via Metadata-guided Diffusion Model
Shuyun Wang, Hu Zhang, Xin Shen, Dadong Wang, Xin Yu; Proceedings of the Compute r Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22975-22984
Bitstream-corrupted video recovery aims to fill in realistic video content due t o bitstream corruption during video storage or transmission. Most existing metho ds typically assume that the predefined masks of the corrupted regions are known in advance. However, manually annotating these masks is laborious and time-cons

uming, limiting the applicability of existing methods in real-world scenarios. T
herefore, we expect to relax this assumption by defining a new blind video recov
ery setting where the recovery of corrupted regions does not rely on predefined
masks. There are two significant challenges in this setting: (i) without predefi
ned masks, how accurately can a model identify the regions requiring recovery? (
ii) how to recover contents from extensive and irregular regions, especially whe
n large portions of frames are severely degraded? To address these challenges, w
e introduce a Metadata-Guided Diffusion Model, dubbed M-GDM. To enable a diffusi
on model focusing on the corrupted regions, we leverage intrinsic video metadata
as a corruption indicator and design a dual-stream metadata encoder. This encod
er first embeds the motion vectors and frame types of a video separately and the
n merges them into a unified metadata representation. The metadata representatio
n will interact with the corrupted latent feature through cross-attention mechan
isms at each diffusion step. Meanwhile, to preserve the intact regions, we propo
se a prior-driven mask predictor that generates pseudo masks for the corrupted r
egions by leveraging the metadata prior and diffusion prior. These pseudo masks
enable the separation and recombination of intact and recovered regions through
hard masking. However, imperfections in pseudo mask predictions and hard masking
processes often result in boundary artifacts. Thus, we introduce a post-refinem
ent module that refines the hard-masked outputs, enhancing the consistency betwe
en intact and recovered regions. Extensive experiment results validate the effec
tiveness of our method and demonstrate its superiority in the blind video recove
ry task.
****

Mind the Trojan Horse: Image Prompt Adapter Enabling Scalable and Deceptive Jail
breaking
Junxi Chen, Junhao Dong, Xiaohua Xie; Proceedings of the Computer Vision and Pat
tern Recognition Conference (CVPR), 2025, pp. 23785-23794
Recently, the Image Prompt Adapter (IP-Adapter) has been increasingly integrated
into text-to-image diffusion models (T2I-DMs) to improve controllability. Howev
er, in this paper, we reveal that T2I-DMs equipped with the IP-Adapter (T2I-IP-D
Ms) enable a new jailbreak attack named the hijacking attack. We demonstrate tha
t, by uploading imperceptible image-space adversarial examples (AEs), the advers
ary can hijack massive benign users to jailbreak an Image Generation Service (IG
S) driven by T2I-IP-DMs and mislead the public to discredit the service provider
. Worse still, the IP-Adapter's dependency on open-source image encoders reduces
the knowledge required to craft AEs. Extensive experiments verify the technical
feasibility of the hijacking attack. In light of the revealed threat, we invest
igate several existing defenses and explore combining the IP-Adapter with advers
arially trained models to overcome existing defenses' limitations. Our code is a
vailable at https://github.com/fhdnskfbeuv/attackIPA.
****

Lost in Translation, Found in Context: Sign Language Translation with Contextual
Cues
Youngjoon Jang, Haran Raajesh, Liliane Momeni, Gül Varol, Andrew Zisserman; Proc
eedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025,
pp. 8742-8752
Our objective is to translate continuous sign language into spoken language text
. Inspired by the way human interpreters rely on context for accurate translatio
n, we incorporate additional contextual cues together with the signing video, in
to a new translation framework. Specifically, besides visual sign recognition fe
atures that encode the input video, we integrate complementary textual informati
on from (i) captions describing the background show, (ii) translation of previou
s sentences, as well as (iii) pseudo-glosses transcribing the signing. These are
automatically extracted and inputted along with the visual features to a pre-tr
ained large language model (LLM), which we fine-tune to generate spoken language
translations in text form. Through extensive ablation studies, we show the posi
tive contribution of each input cue to the translation performance. We train and
evaluate our approach on BOBSL -- the largest British Sign Language dataset cur
rently available. We show that our contextual approach significantly enhances th

e quality of the translations compared to previously reported results on BOBSL, and also to state-of-the-art methods that we implement as baselines. Furthermore, we demonstrate the generality of our approach by applying it also to How2Sign, an American Sign Language dataset, and achieve competitive results.
****

CoCoGaussian: Leveraging Circle of Confusion for Gaussian Splatting from Defocused Images
Jungho Lee, Suhwan Cho, Taeoh Kim, Ho-Deok Jang, Minhyeok Lee, Geonho Cha, Dongyoon Wee, Dogyoon Lee, Sangyoun Lee; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16101-16110
3D Gaussian Splatting (3DGS) has attracted significant attention for its high-quality novel view rendering, inspiring research to address real-world challenges. While conventional methods depend on sharp images for accurate scene reconstruction, real-world scenarios are often affected by defocus blur due to finite depth of field, making it essential to account for realistic 3D scene representation. In this study, we propose CoCoGaussian, a Circle of Confusion-aware Gaussian Splatting that enables precise 3D scene representation using only defocused images. CoCoGaussian addresses the challenge of defocus blur by modeling the Circle of Confusion (CoC) through a physically grounded approach based on the principles of photographic defocus. Exploiting 3D Gaussians, we compute the CoC diameter from depth and learnable aperture information, generating multiple Gaussians to precisely capture the CoC shape. Furthermore, we introduce a learnable scaling factor to enhance robustness and provide more flexibility in handling unreliable depth in scenes with reflective or refractive surfaces. Experiments on both synthetic and real-world datasets demonstrate that CoCoGaussian achieves state-of-the-art performance across multiple benchmarks.
****

Semantic and Sequential Alignment for Referring Video Object Segmentation
Feiyu Pan, Hao Fang, Fangkai Li, Yanyu Xu, Yawei Li, Luca Benini, Xiankai Lu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19067-19076
Referring video object segmentation (RVOS) seeks to segment the objects within a video referred by linguistic expressions. Existing RVOS solutions follow a "fuse then select" paradigm: establishing semantic correlation between visual and linguistic feature, and performing frame-level query interaction to select the instance mask per frame with instance segmentation module. This paradigm overlooks the challenge of semantic gap between the linguistic descriptor and the video object as well as the underlying clutters in the video. This paper proposes a novel Semantic and Sequential Alignment (SSA) paradigm to handle these challenges. We first insert a lightweight adapter after the vision language model (VLM) to perform the semantic alignment. Then, prior to selecting mask per frame, we exploit the trajectory-to-instance enhancement for each frame via sequential alignment. This paradigm leverages the visual-language alignment inherent in VLM during adaptation and tries to capture global information by ensembling trajectories. This helps understand videos and the corresponding descriptors by mitigating the discrepancy with intricate activity semantics, particularly when facing occlusion or similar interference. SSA demonstrates competitive performance while maintaining fewer learnable parameters.
****

Synchronized Video-to-Audio Generation via Mel Quantization-Continuum Decomposition
Juncheng Wang, Chao Xu, Cheng Yu, Lei Shang, Zhe Hu, Shujun Wang, Liefeng Bo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3111-3120
Video-to-audio generation is essential for synthesizing realistic audio tracks that synchronize effectively with silent videos.Following the perspective of extracting essential signals from videos that can precisely control the mature text-to-audio generative diffusion models, this paper presents how to balance the representation of mel-spectrograms in terms of completeness and complexity through a new approach called Mel Quantization-Continuum Decomposition (Mel-QCD).We deco

mpose the mel-spectrogram into three distinct types of signals, employing quanti zation or continuity to them, we can effectively predict them from video by a de vised video-to-all (V2X) predictor.Then, the predicted signals are recomposed an d fed into a ControlNet, along with a textual inversion design, to control the a udio generation process.Our proposed Mel-QCD method demonstrates state-of-the-ar t performance across eight metrics, evaluating dimensions such as quality, synch ronization, and semantic consistency.
****

Continual SFT Matches Multimodal RLHF with Negative Supervision
Ke Zhu, Yu Wang, Yanpeng Sun, Qiang Chen, Jiangjiang Liu, Gang Zhang, Jingdong W ang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14615-14624
Multimodal RLHF usually happens after supervised finetuning (SFT) stage to conti nually improve vision-language models' (VLMs) comprehension. Conventional wisdom holds its superiority over continual SFT during this preference alignment stage . In this paper, we observe that the inherent value of multimodal RLHF lies in i ts negative supervision, the logit of the rejected responses. We thus propose a novel negative supervised finetuning (nSFT) approach that fully excavates these information resided. Our nSFT disentangles this negative supervision in RLHF par adigm, and continually aligns VLMs with a simple SFT loss. This is more memory e fficient than multimodal RLHF where 2 (e.g., DPO) or 4 (e.g., PPO) large VLMs ar e strictly required. The effectiveness of nSFT is rigorously proved by comparing it with various multimodal RLHF approaches, across different dataset sources, b ase VLMs and evaluation metrics. Besides, fruitful of ablations are provided to support our hypothesis. Code will be found in https://github.com/Kevinz-code/nSF T/.
****

Semantic-guided Cross-Modal Prompt Learning for Skeleton-based Zero-shot Action Recognition
Anqi Zhu, Jingmin Zhu, James Bailey, Mingming Gong, Qiuhong Ke; Proceedings of t he Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13876-13 885
Skeleton-based human action recognition is promising due to its privacy preserva tion, robustness to visual challenges, and computational efficiency. Especially, the practical necessity to recognize unseen actions has led to increased intere st in zero-shot skeleton-based action recognition (ZSSAR). Existing ZSSAR approa ches often rely on manually crafted action descriptions or visual assumptions to enhance knowledge transfer, which is limited in flexibility and prone to inaccu racies and noise. To overcome this, we introduce Semantic-guided Cross-Modal Pro mpt Learning (SCoPLe), a novel framework that replaces manual guidance with data -driven prompt learning for refinement and alignment of skeletal and textual fe atures. Specifically, we introduce a dual-stream language prompting module that preserves the original semantic context from the pre-trained text encoder while still effectively tuning its ouput for ZSSAR task adaptation. We also introduce a joint-shaped prompting module that learns tuning for skeleton features and inc orporate an adaptive visual representation sampler that leverages text semantics to strengthen the cross-modal prompting interactions during skeleton-to-text em bedding projection. Experimental results on the NTU-RGB+D and PKU-MMD datasets d emonstrate the state-of-the-art performance of our method in both ZSSAR and gene ralized ZSSAR scenarios.
****

FATE: Full-head Gaussian Avatar with Textural Editing from Monocular Video
Jiawei Zhang, Zijian Wu, Zhiyang Liang, Yicheng Gong, Dongfang Hu, Yao Yao, Xun Cao, Hao Zhu; Proceedings of the Computer Vision and Pattern Recognition Confere nce (CVPR), 2025, pp. 5535-5545
Reconstructing high-fidelity, animatable 3D head avatars from effortlessly captu red monocular videos is a pivotal yet formidable challenge. Although significant progress has been made in rendering performance and manipulation capabilities, notable challenges remain, including incomplete reconstruction and inefficient G aussian representation. To address these challenges, we introduce FATE -- a nove

l method for reconstructing an editable full-head avatar from a single monocular video. FATE integrates a sampling-based densification strategy to ensure optimal positional distribution of points, improving rendering efficiency. A neural baking technique is introduced to convert discrete Gaussian representations into continuous attribute maps, facilitating intuitive appearance editing. Furthermore, we propose a universal completion framework to recover non-frontal appearance, culminating in a 360^\circ-renderable 3D head avatar. FATE outperforms previous approaches in both qualitative and quantitative evaluations, achieving state-of-the-art performance. To the best of our knowledge, FATE is the first animatable and 360^\circ full-head monocular reconstruction method for a 3D head avatar. Project page and code are available at this \href https://zjwfufu.github.io/FATE-page/ link .

****

ChatGen: Automatic Text-to-Image Generation From FreeStyle Chatting

Chengyou Jia, Changliang Xia, Zhuohang Dang, Weijia Wu, Hangwei Qian, Minnan Luo ; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13284-13293

Despite the significant advancements in text-to-image (T2I) generative models, users often face a trial-and-error challenge in practical scenarios. This challenge arises from the complexity and uncertainty of tedious steps such as crafting suitable prompts, selecting appropriate models, and configuring specific arguments, making users resort to labor-intensive attempts for desired images. This paper proposes Automatic T2I generation, which aims to automate these tedious steps, allowing users to simply describe their needs in a freestyle chatting way. To systematically study this problem, we first introduce ChatGenBench, a novel benchmark designed for Automatic T2I. It features high-quality paired data with diverse freestyle inputs, enabling comprehensive evaluation of automatic T2I models across all steps. Additionally, recognizing Automatic T2I as a complex multi-step reasoning task, we propose ChatGen-Evo, a multi-stage evolution strategy that progressively equips models with essential automation skills. Through extensive evaluation across step-wise accuracy and image quality, ChatGen-Evo significantly enhances performance over various baselines. Our evaluation also uncovers valuable insights for advancing automatic T2I. All our data, code and models will be publicly available.

****

GEM: A Generalizable Ego-Vision Multimodal World Model for Fine-Grained Ego-Motion, Object Dynamics, and Scene Composition Control

Mariam Hassan, Sebastian Stapf, Ahmad Rahimi, Pedro M B Rezende, Yasaman Haghighi, David Brüggemann, Isinsu Katircioglu, Lin Zhang, Xiaoran Chen, Suman Saha, Marco Cannici, Elie Aljalbout, Botao Ye, Xi Wang, Aram Davtyan, Mathieu Salzmann, Davide Scaramuzza, Marc Pollefeys, Paolo Favaro, Alexandre Alahi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22404-22415

We present GEM, a Generalizable Ego-vision Multimodal world model that predicts future frames using a reference frame, sparse features, human poses, and ego-trajectories. Hence, our model has precise control over object dynamics, ego-agent motion and human poses. GEM generates paired RGB and depth outputs for richer spatial understanding. We introduce autoregressive noise schedules to enable stable long-horizon generations. Our dataset is comprised of 4000+ hours of multimodal data across domains like autonomous driving, egocentric human activities, and drone flights. Pseudo-labels are used to get depth maps, ego-trajectories, and human poses. We use a comprehensive evaluation framework, including a new Control of Object Manipulation (COM) metric, to assess controllability. Experiments show GEM excels at generating diverse, controllable scenarios and temporal consistency over long generations. Code, models, and datasets are fully open-sourced.

****

VEU-Bench: Towards Comprehensive Understanding of Video Editing

Bozheng Li, Yongliang Wu, Yi Lu, Jiashuo Yu, Licheng Tang, Jiawang Cao, Wenqing Zhu, Yuyang Sun, Jay Wu, Wenbo Zhu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13671-13680

Widely shared videos on the internet are often edited. Recently, although Video Large Language Models (Vid-LLMs) have made great progress in general video understanding tasks, their capabilities in video editing understanding (VEU) tasks remain unexplored. To address this gap, in this paper, we introduce VEU-Bench (Video Editing Understanding Benchmark), a comprehensive benchmark that categorizes video editing components across various dimensions, from intra-frame features like shot size to inter-shot attributes such as cut types and transitions. Unlike previous video editing understanding benchmarks that focus mainly on editing element classification, VEU-Bench encompasses 19 fine-grained tasks across three stages: recognition, reasoning, and judging. To enhance the annotation of VEU automatically, we built an annotation pipeline integrated with an ontology-based knowledge base. Through extensive experiments with 11 state-of-the-art Vid-LLMs, our findings reveal that current Vid-LLMs face significant challenges in VEU tasks, with some performing worse than random choice. To alleviate this issue, we develop Oscars(Named after the Academy Awards.), a VEU expert model fine-tuned on the curated VEU-Bench dataset. It outperforms existing open-source Vid-LLMs on VEU-Bench by over 28.3% in accuracy and achieves performance comparable to commercial models like GPT-4o. We also demonstrate that incorporating VEU data significantly enhances the performance of Vid-LLMs on general video understanding benchmarks, with an average improvement of 8.3% across nine reasoning tasks. The code and data will be made available.

****

Decouple Distortion from Perception: Region Adaptive Diffusion for Extreme-low Bitrate Perception Image Compression

Jinchang Xu, Shaokang Wang, Jintao Chen, Zhe Li, Peidong Jia, Fei Zhao, Guoqing Xiang, Zhijian Hao, Shanghang Zhang, Xiaodong Xie; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18051-18061

Leveraging the generative power of diffusion models, generative image compression has achieved impressive perceptual fidelity even at extremely low bitrates. However, current methods often neglect the non-uniform complexity of images, limiting their ability to balance global perceptual quality with local texture consistency and to allocate coding resources efficiently. To address this, we introduce the Map-guided Masking Realism Image Diffusion Codec(MRIDC), designed to optimize the trade-off between local distortion and global perceptual quality in extreme-low bitrate compression. MRIDC integrates a vector-quantized image encoder with a diffusion-based decoder. On the encoding side, we propose a Map-guided Latent Masking(MLM) module, which selectively masks elements in the latent space based on prior information, allowing adaptive resource allocation aligned with image complexity. On the decoding side, masked latents are completed using the Bidirectional Prediction Controllable Generation(BPCG) module, which guides the constrained generation process within the diffusion model to reconstruct the image. Experimental results show that MRIDC achieves state-of-the-art perceptual compression quality at extremely low bitrates, effectively preserving feature consistency in key regions and advancing the rate-distortion-perception performance curve, establishing new benchmarks in balancing compression efficiency with visual fidelity.

****

Yo'Chameleon: Personalized Vision and Language Generation

Thao Nguyen, Krishna Kumar Singh, Jing Shi, Trung Bui, Yong Jae Lee, Yuheng Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14438-14448

Large Multimodal Models (e.g., GPT-4, Gemini, Chameleon) have evolved into powerful tools with millions of users. However, they remain generic models and lack personalized knowledge of specific user concepts. Previous work has explored personalization for text generation, yet it remains unclear how these methods can be adapted to new modalities, such as image generation. In this paper, we introduce Yo'Chameleon, the first attempt to study personalization for large multimodal models. Given 3-5 images of a particular concept, Yo'Chameleon leverages soft-prompt tuning to embed subject-specific information to (i) answer questions about the subject and (ii) recreate pixel-level details to produce images of the subje

ct in new contexts. Yo'Chameleon is trained with (i) a self-prompting optimizati on mechanism to balance performance across multiple modalities, and (ii) a "soft -positive" image generation approach to enhance image quality in a few-shot sett ing. Our qualitative and quantitative analyses reveal that Yo'Chameleon can lear n concepts more efficiently using fewer tokens and effectively encode visual att ributes, outperforming prompting baselines.
****

PatchVSR: Breaking Video Diffusion Resolution Limits with Patch-wise Video Super -Resolution
Shian Du, Menghan Xia, Chang Liu, Xintao Wang, Jing Wang, Pengfei Wan, Di Zhang, Xiangyang Ji; Proceedings of the Computer Vision and Pattern Recognition Confer ence (CVPR), 2025, pp. 17799-17809
Pre-trained video generation models hold great potential for generative video su per-resolution (VSR). However, adapting them for full-size VSR, as most existing methods do, suffers from unnecessary intensive full-attention computation and f ixed output resolution. To overcome these limitations, we make the first explora tion into utilizing video diffusion priors for patch-wise VSR.This is non-trivia l because pre-trained video diffusion models are not native for patch-level deta il generation. To mitigate this challenge, we propose an innovative approach, ca lled PatchVSR, which integrates a dual-stream adapter for conditional guidance. The patch branch extracts features from input patches to maintain content fideli ty while the global branch extracts context features from the resized full video to bridge the generation gap caused byincomplete semantics of patches.Particula rly, we also inject the patch's location information into the model to better co ntextualize patch synthesis within the global video frame.Experiments demonstrat e that our method can synthesize high-fidelity, high-resolution details at the p atch level. A tailor-made multi-patch joint modulation is proposed to ensure vis ual consistency across individually enhanced patches. Due to the flexibility of our patch-based paradigm, we can achieve highly competitive 4K VSR based on a 51 2x512 resolution base model, with extremely high efficiency.
****

FluxSpace: Disentangled Semantic Editing in Rectified Flow Models
Yusuf Dalva, Kavana Venkatesh, Pinar Yanardag; Proceedings of the Computer Visio n and Pattern Recognition Conference (CVPR), 2025, pp. 13083-13092
Rectified flow models have emerged as a dominant approach in image generation, s howcasing impressive capabilities in high-quality image synthesis. However, desp ite their effectiveness in visual generation, rectified flow models often strugg le with disentangled editing of images. This limitation prevents the ability to perform precise, attribute-specific modifications without affecting unrelated as pects of the image. In this paper, we introduce FluxSpace, a domain-agnostic ima ge editing method leveraging a representation space with the ability to control the semantics of images generated by rectified flow transformers, such as Flux. By leveraging the representations learned by the transformer blocks within the r ectified flow models, we propose a set of semantically interpretable representat ions that enable a wide range of image editing tasks, from fine-grained image ed iting to artistic creation. This work offers a scalable and effective image edit ing approach, along with its disentanglement capabilities.
****

Scene-Centric Unsupervised Panoptic Segmentation
Oliver Hahn, Christoph Reich, Nikita Araslanov, Daniel Cremers, Christian Ruppre cht, Stefan Roth; Proceedings of the Computer Vision and Pattern Recognition Con ference (CVPR), 2025, pp. 24485-24495
Unsupervised panoptic segmentation aims to partition an image into semantically meaningful regions and distinct object instances without training on manually an notated data. In contrast to prior work on unsupervised panoptic scene understan ding, we eliminate the need for object-centric training data, enabling the unsup ervised understanding of complex scenes. To that end, we present the first unsu pervised panoptic method that directly trains on scene-centric imagery. In parti cular, we propose an approach to obtain high-resolution panoptic pseudo labels o n complex scene-centric data, combining visual representations, depth, and motio

n cues. Utilizing both pseudo-label training and a panoptic self-training strategy yields a novel approach that accurately predicts panoptic segmentation of complex scenes without requiring any human annotations. Our approach significantly improves panoptic quality, e.g., surpassing the recent state of the art in unsupervised panoptic segmentation on Cityscapes by 9.4% points in PQ.
****

Touch2Shape: Touch-Conditioned 3D Diffusion for Shape Exploration and Reconstruction

Yuanbo Wang, Zhaoxuan Zhang, Jiajin Qiu, Dilong Sun, Zhengyu Meng, Xiaopeng Wei, Xin Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5656-5665

Diffusion models have made breakthroughs in 3D generation tasks. Current 3D diffusion models focus on reconstructing target shape from images or a set of partial observations. While excelling in global context understanding, they struggle to capture the local details of complex shapes and limited to the occlusion and lighting conditions. To overcome these limitations, we utilize tactile images to capture the local 3D information and propose a Touch2Shape model, which leverages a touch-conditioned diffusion model to explore and reconstruct the target shape from touch. For shape reconstruction, we have developed a touch embedding module to condition the diffusion model in creating a compact representation and a touch shape fusion module to refine the reconstructed shape. For shape exploration, we combine the diffusion model with reinforcement learning to train a policy. This involves using the generated latent vector from the diffusion model to guide the touch exploration policy training through a novel reward design. Experiments validate the reconstruction quality thorough both qualitatively and quantitative analysis, and our touch exploration policy further boosts reconstruction performance.
****

VITED: Video Temporal Evidence Distillation

Yujie Lu, Yale Song, William Wang, Lorenzo Torresani, Tushar Nagarajan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8501-8511

We investigate complex video question answering via chain-of-evidence reasoning --- identifying sequences of temporal spans from multiple relevant parts of the video, together with visual evidence within them.Existing models struggle with multi-step reasoning as they uniformly sample a fixed number of frames, which can miss critical evidence distributed nonuniformly throughout the video. Moreover, they lack the ability to temporally localize such evidence in the broader context of the full video, which is required for answering complex questions. We propose a framework to enhance existing VideoQA datasets with evidence reasoning chains, automatically constructed by searching for optimal intervals of interest in the video with supporting evidence, that maximizes the likelihood of answering a given question.We train our model (ViTED) to generate these evidence chains directly, enabling it to both localize evidence windows as well as perform multi-step reasoning across them in long-form video content.We show the value of our evidence-distilled models on a suite of long video QA benchmarks where we outperform state-of-the-art approaches that lack evidence reasoning capabilities.
****

Adversarial Domain Prompt Tuning and Generation for Single Domain Generalization

Zhipeng Xu, De Cheng, Xinyang Jiang, Nannan Wang, Dongsheng Li, Xinbo Gao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18584-18595

Single domain generalization (SDG) aims to learn a robust model, which could perform well on many unseen domains while there is only one single domain available for training. One of the promising directions for achieving single-domain generalization is to generate out-of-domain (OOD) training data through data augmentation or image generation. Given the rapid advancements in AI-generated content (AIGC), this paper is the first to propose leveraging powerful pre-trained text-to-image (T2I) foundation models to create the training data. However, manually designing textual prompts to generate images for all possible domains is often im

practical, and some domain characteristics may be too abstract to describe with words. To address these challenges, we propose a novel Progressive Adversarial Prompt Tuning (PAPT) framework for pre-trained diffusion models. Instead of relying on static textual domains, our approach learns two sets of abstract prompts as conditions for the diffusion model: one that captures domain-invariant category information and another that models domain-specific styles. This adversarial learning mechanism enables the T2I model to generate images in various domain styles while preserving key categorical features. Extensive experiments demonstrate the effectiveness of the proposed method, achieving superior performances to state-of-the-art single-domain generalization approaches.
****

Learning Physics From Video: Unsupervised Physical Parameter Estimation for Continuous Dynamical Systems
Alejandro Castañeda Garcia, Jan Warchocki, Jan van Gemert, Daan Brinks, Nergis Tomen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27924-27933
Extracting physical dynamical system parameters from recorded observations is key in natural science. Current methods for automatic parameter estimation from video train supervised deep networks on large datasets. Such datasets require labels, which are difficult to acquire. While some unsupervised techniques--which depend on frame prediction--exist, they suffer from long training times, initialization instabilities, only consider motion-based dynamical systems, and are evaluated mainly on synthetic data. In this work, we propose an unsupervised method to estimate the physical parameters of known, continuous governing equations from single videos suitable for different dynamical systems beyond motion and robust to initialization. Moreover, we remove the need for frame prediction by implementing a KL-divergence-based loss function in the latent space, which avoids convergence to trivial solutions and reduces model size and compute. We first evaluate our model on synthetic data, as commonly done. After which, we take the field closer to reality by recording Delfys75: our own real-world dataset of 75 videos for five different types of dynamical systems to evaluate our method and others. Our method compares favorably to others. Code and data are available online: https://github.com/Alejandro-neuro/Learning_physics_from_video.
****

Temporal Score Analysis for Understanding and Correcting Diffusion Artifacts
Yu Cao, Zengqun Zhao, Ioannis Patras, Shaogang Gong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7707-7716
Visual artifacts remain a persistent challenge in diffusion models, even with training on massive datasets. Current solutions primarily rely on supervised detectors, yet lack understanding of why these artifacts occur in the first place. In our analysis, we identify three distinct phases in the diffusion generative process: Profiling, Mutation, and Refinement. Artifacts typically emerge during the Mutation phase, where certain regions exhibit anomalous score dynamics over time, causing abrupt disruptions in the normal evolution pattern. This temporal nature explains why existing methods focusing only on spatial uncertainty of the final output fail at effective artifact localization. Based on these insights, we propose ASCED (Abnormal Score Correction for Enhancing Diffusion), that detects artifacts by monitoring abnormal score dynamics during the diffusion process, with a trajectory-aware on-the-fly mitigation strategy that appropriate generation of noise in the detected areas. Unlike most existing methods that apply post hoc corrections, e.g., by applying a noising-denoising scheme after generation, our mitigation strategy operates seamlessly within the existing diffusion process. Extensive experiments demonstrate that our proposed approach effectively reduces artifacts across diverse domains, matching or surpassing existing supervised methods without additional training.
****

ProAPO: Progressively Automatic Prompt Optimization for Visual Classification
Xiangyan Qu, Gaopeng Gou, Jiamin Zhuang, Jing Yu, Kun Song, Qihao Wang, Yili Li, Gang Xiong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25145-25155

Vision-language models (VLMs) have made significant progress in image classification by training with large-scale paired image-text data. Their performances largely depend on the prompt quality. While recent methods show that visual descriptions generated by large language models (LLMs) enhance the generalization of VLMs, class-specific prompts may be inaccurate or lack discrimination due to the hallucination in LLMs. In this paper, we aim to find visually discriminative prompts for fine-grained categories with minimal supervision and no human-in-the-loop. An evolution-based algorithm is proposed to progressively optimize language prompts from task-specific templates to class-specific descriptions. Unlike optimizing templates, the search space shows an explosion in class-specific candidate prompts. This increases prompt generation costs, iterative times, and the overfitting problem. To this end, we first introduce several simple yet effective edit-based and evolution-based operations to generate diverse candidate prompts by one-time query of LLMs. Then, two sampling strategies are proposed to find a better initial search point and reduce traversed categories, saving iteration costs. Moreover, we apply a novel fitness score with entropy constraints to mitigate overfitting. In a challenging one-shot image classification setting, our method outperforms existing textual prompt-based methods and improves LLM-generated description methods across 13 datasets. Meanwhile, we demonstrate that our optimal prompts improve adapter-based methods and transfer effectively across different backbones.

****

ShapeWords: Guiding Text-to-Image Synthesis with 3D Shape-Aware Prompts

Dmitry Petrov, Pradyumn Goyal, Divyansh Shivashok, Yuanming Tao, Melinos Averkiou, Evangelos Kalogerakis; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13305-13314

We introduce ShapeWords, an approach for synthesizing images based on 3D shape guidance and text prompts.ShapeWords incorporates target 3D shape information within specialized tokens embedded together with the input text, effectively blending 3D shape awareness with textual context to guide the image synthesis process. Unlike conventional shape guidance methods that rely on depth maps restricted to fixed viewpoints and often overlook full 3D structure or textual context, ShapeWords generates diverse yet consistent images that reflect both the target shape's geometry and the textual description. Experimental results show that ShapeWords produces images that are more text-compliant, aesthetically plausible, while also maintaining 3D shape awareness.

****

Auto-Encoded Supervision for Perceptual Image Super-Resolution

MinKyu Lee, Sangeek Hyun, Woojin Jun, Jae-Pil Heo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17958-17968

This work tackles the fidelity objective in the perceptual super-resolution (SR) task. Specifically, we address the shortcomings of pixel-level $\mathcal L _\text p$ loss ($\mathcal L _\text pix$) in the GAN-based SR framework. Since $\mathcal L _\text pix$ is known to have a trade-off relationship against perceptual quality, prior methods often multiply a small scale factor or utilize low-pass filters. However, this work shows that these circumventions fail to address the fundamental factor that induces blurring. Accordingly, we focus on two points: 1) precisely discriminating the subcomponent of $\mathcal L _\text pix$ that contributes to blurring, and 2) guiding reconstruction only based on the factor that is free from this trade-off relationship. We show that this can be achieved in a surprisingly simple manner, with an Auto-Encoder (AE) pretrained using $\mathcal L _\text pix$. Based on this insight, we propose the Auto-Encoded Supervision for Optimal Penalization loss ($\mathcal L _\text AESOP$), a novel loss function that measures distance in the AE space (the space after the decoder, not the bottleneck), rather than in the raw pixel space. By simply substituting $\mathcal L _\text pix$ with $\mathcal L _\text AESOP$, we can provide effective reconstruction guidance without compromising perceptual quality. Designed for simplicity, our method enables easy integration into existing SR frameworks. Extensive experiments demonstrate the effectiveness of AESOP.

****

Black Swan: Abductive and Defeasible Video Reasoning in Unpredictable Events
Aditya Chinchure, Sahithya Ravi, Raymond Ng, Vered Shwartz, Boyang Li, Leonid Sigal; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24201-24210
The commonsense reasoning capabilities of vision-language models (VLMs), especially in abductive reasoning and defeasible reasoning, remain poorly understood. Most benchmarks focus on typical visual scenarios, making it difficult to discern whether model performance stems from keen perception and reasoning skills, or reliance on pure statistical recall. We argue that by focusing on atypical events in videos, clearer insights can be gained on the core capabilities of VLMs. Explaining and understanding such out-of-distribution events requires models to extend beyond basic pattern recognition and regurgitation of their prior knowledge. To this end, we introduce BlackSwanSuite, a benchmark for evaluating VLMs' ability to reason about unexpected events through abductive and defeasible tasks. Our tasks artificially limit the amount of visual information provided to models while questioning them about hidden unexpected events, or provide new visual information that could change an existing hypothesis about the event. We curate a comprehensive benchmark suite comprising over 3,800 MCQ, 4,900 generative and 6,700 yes/no questions, spanning 1,655 videos. After extensively evaluating various state-of-the-art VLMs, including GPT-4o and Gemini 1.5 Pro, as well as open-source VLMs such as LLaVA-Video, we find significant performance gaps of up to 32% from humans on these tasks. Our findings reveal key limitations in current VLMs, emphasizing the need for enhanced model architectures and training strategies. Our data and leaderboard is available at blackswan.cs.ubc.ca.
****
Go-with-the-Flow: Motion-Controllable Video Diffusion Models Using Real-Time Warped Noise
Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, Michael Ryoo, Paul Debevec, Ning Yu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13-23
Generative modeling aims to transform random noise into structured outputs. In this work, we enhance video diffusion models by allowing motion control via structured latent noise sampling. This is achieved by just a change in data: we pre-process training videos to yield structured noise. Consequently, our method is agnostic to diffusion model design, requiring no changes to model architectures or training pipelines. Specifically, we propose a novel noise warping algorithm, fast enough to run in real time, that replaces random temporal Gaussianity with correlated warped noise derived from optical flow fields, while preserving the spatial Gaussianity. The efficiency of our algorithm enables us to fine-tune modern video diffusion base models using warped noise with minimal overhead, and provide a one-stop solution for a wide range of user-friendly motion control: local object motion control, global camera movement control, and motion transfer. The harmonization between temporal coherence and spatial Gaussianity in our warped noise leads to effective motion control while maintaining per-frame pixel quality. Extensive experiments and user studies demonstrate the advantages of our method, making it a robust and scalable approach for controlling motion in video diffusion models. Please see our project webpage: https://eyeline-research.github.io/Go-with-the-Flow/ ; source code and checkpoints are available on GitHub: https://github.com/Eyeline-Research/Go-with-the-Flow
****
Silence is Golden: Leveraging Adversarial Examples to Nullify Audio Control in LDM-based Talking-Head Generation
Yuan Gan, Jiaxu Miao, Yunze Wang, Yi Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13434-13444
Advances in talking-head animation based on Latent Diffusion Models (LDM) enable the creation of highly realistic, synchronized videos. These fabricated videos are indistinguishable from real ones, increasing the risk of potential misuse for scams, political manipulation, and misinformation. Hence, addressing these ethical concerns has become a pressing issue in AI security. Recent proactive defen

se studies focused on countering LDM-based models by adding perturbations to portraits. However, these methods are ineffective at protecting reference portraits from advanced image-to-video animation. The limitations are twofold: 1) they fail to prevent images from being manipulated by audio signals, and 2) diffusion-based purification techniques can effectively eliminate protective perturbations. To address these challenges, we propose Silencer, a two-stage method designed to proactively protect the privacy of portraits. First, a nullifying loss is proposed to ignore audio control in talking-head generation. Second, we apply anti-purification loss in LDM to optimize the inverted latent feature to generate robust perturbations. Extensive experiments demonstrate the effectiveness of Silencer in proactively protecting portrait privacy. We hope this work will raise awareness among the AI security community regarding critical ethical issues related to talking-head generation techniques. Code: https://github.com/yuangan/Silencer.

****

Iterative Predictor-Critic Code Decoding for Real-World Image Dehazing

Jiayi Fu, Siyu Liu, Zikun Liu, Chun-Le Guo, Hyunhee Park, Ruiqi Wu, Guoqing Wang, Chongyi Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12700-12709

We propose a novel Iterative Predictor-Critic Code Decoding framework for real-world image dehazing, abbreviated as IPC-Dehaze, which leverages the high-quality codebook prior encapsulated in a pre-trained VQGAN. Apart from previous codebook-based methods that rely on one-shot decoding, our method utilizes high-quality codes obtained in the previous iteration to guide the prediction of the Code-Predictor in the subsequent iteration, improving code prediction accuracy and ensuring stable dehazing performance. Our idea stems from the observations that 1) the degradation of hazy images varies with haze density and scene depth, and 2) clear regions play crucial cues in restoring dense haze regions. However, it is nontrivial to progressively refine the obtained codes in subsequent iterations, owing to the difficulty in determining which codes should be retained or replaced at each iteration. Another key insight of our study is to propose Code-Critic to capture interrelations among codes. The Code-Critic is used to evaluate code correlations and then resample a set of codes with the highest mask scores, i.e., a higher score indicates that the code is more likely to be rejected, which helps retain more accurate codes and predict difficult ones. Extensive experiments demonstrate the superiority of our method over state-of-the-art methods in real-world dehazing. Our code will be made publicly available.

****

RNG: Relightable Neural Gaussians

Jiahui Fan, Fujun Luan, Jian Yang, Milos Hasan, Beibei Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26525-26534

3D Gaussian Splatting (3DGS) has shown impressive results for the novel view synthesis task, where lighting is assumed to be fixed. However, creating relightable 3D assets, especially for objects with ill-defined shapes (fur, fabric, etc.), remains a challenging task. The decomposition between light, geometry, and material is ambiguous, especially if either smooth surface assumptions or surface-based analytical shading models do not apply. We propose Relightable Neural Gaussians (RNG), a novel 3DGS-based framework that enables the relighting of objects with both hard surfaces or soft boundaries, while avoiding assumptions on the shading model. We condition the radiance at each point on both view and light directions. We also introduce a shadow cue, as well as a depth refinement network to improve shadow accuracy. Finally, we propose a hybrid forward-deferred fitting strategy to balance geometry and appearance quality. Our method achieves significantly faster training (1.3 hours) and rendering (60 frames per second) compared to a prior method based on neural radiance fields and produces higher-quality shadows than a concurrent 3DGS-based method.

****

Towards Realistic Example-based Modeling via 3D Gaussian Stitching

Xinyu Gao, Ziyi Yang, Bingchen Gong, Xiaoguang Han, Sipeng Yang, Xiaogang Jin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26597-26607

Using parts of existing models to rebuild new models, commonly termed as example-based modeling, is a classical methodology in the realm of computer graphics. Previous works mostly focus on shape composition, making them very hard to use for realistic composition of 3D objects captured from real-world scenes. This leads to combining multiple NeRFs into a single 3D scene to achieve seamless appearance blending. However, the current SeamlessNeRF method struggles to achieve interactive editing and harmonious stitching for real-world scenes due to its gradient-based strategy and grid-based representation. To this end, we present an example-based modeling method that combines multiple Gaussian fields in a point-based representation using sample-guided synthesis. Specifically, as for composition, we create a GUI to segment and transform multiple fields in real time, easily obtaining a semantically meaningful composition of models represented by 3D Gaussian Splatting (3DGS). For texture blending, due to the discrete and irregular nature of 3DGS, straightforwardly applying gradient propagation as SeamlssNeRF is not supported. Thus, a novel sampling-based cloning method is proposed to harmonize the blending while preserving the original rich texture and content. Our workflow consists of three steps: 1) real-time segmentation and transformation of a Gaussian model using a well-tailored GUI, 2) KNN analysis to identify boundary points in the intersecting area between the source and target models, and 3) two-phase optimization of the target model using sampling-based cloning and gradient constraints. Extensive experimental results validate that our approach significantly outperforms previous works in terms of realistic synthesis, demonstrating its practicality.

****

Filter Images First, Generate Instructions Later: Pre-Instruction Data Selection for Visual Instruction Tuning

Bardia Safaei, Faizan Siddiqui, Jiacong Xu, Vishal M. Patel, Shao-Yuan Lo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14247-14256

Visual instruction tuning (VIT) for large vision-language models (LVLMs) requires training on expansive datasets of image-instruction pairs, which can be costly. Recent efforts in VIT data selection aim to select a small subset of high-quality image-instruction pairs, reducing VIT runtime while maintaining performance comparable to full-scale training. However, a major challenge often overlooked is that generating instructions from unlabeled images for VIT is highly expensive. Most existing VIT datasets rely heavily on human annotations or paid services like the GPT API, which limits users with constrained resources from creating VIT datasets for custom applications. To address this, we introduce Pre-Instruction Data Selection (PreSel), a more practical data selection paradigm that directly selects the most beneficial unlabeled images and generates instructions only for the selected images. PreSel first estimates the relative importance of each vision task within VIT datasets to derive task-wise sampling budgets. It then clusters image features within each task, selecting the most representative images with the budget. This approach reduces computational overhead for both instruction generation during VIT data formation and LVLM fine-tuning. By generating instructions for only 15% of the images, PreSel achieves performance comparable to full-data VIT on the LLaVA-1.5 and Vision-Flan datasets. The link to our project page: https://bardisafa.github.io/PreSel

****

Gradient-Guided Annealing for Domain Generalization

Aristotelis Ballas, Christos Diou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20558-20568

Domain Generalization (DG) research has gained considerable traction as of late, since the ability to generalize to unseen data distributions is a requirement that eludes even state-of-the-art training algorithms. In this paper we observe that the initial iterations of model training play a key role in domain generalization effectiveness, since the loss landscape may be significantly different across the training and test distributions, contrary to the case of i.i.d. data. Conflicts between gradients of the loss components of each domain lead the optimization procedure to undesirable local minima that do not capture the domain-invar

iant features of the target classes. We propose alleviating domain conflicts in model optimization, by iteratively annealing the parameters of a model in the ea rly stages of training and searching for points where gradients align between do mains. By discovering a set of parameter values where gradients are updated towa rds the same direction for each data distribution present in the training set, t he proposed Gradient-Guided Annealing (GGA) algorithm encourages models to seek out minima that exhibit improved robustness against domain shifts. The efficacy of GGA is evaluated on five widely accepted and challenging image classification domain generalization benchmarks, where its use alone is able to establish high ly competitive or even state-of-the-art performance. Moreover, when combined wit h previously proposed domain-generalization algorithms it is able to consistentl y improve their effectiveness by significant margins.

****

Generative Sparse-View Gaussian Splatting

Hanyang Kong, Xingyi Yang, Xinchao Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26745-26755

Novel view synthesis from limited observations remains a significant challenge d ue to the lack of information in under-sampled regions, often resulting in notic eable artifacts. We introduce Generative Sparse-view Gaussian Splatting (GS-GS), a general pipeline designed to enhance the rendering quality of 3D/4D Gaussian Splatting (GS) when training views are sparse. Our method generates unseen views using generative models, specifically leveraging pre-trained image diffusion mo dels to iteratively refine view consistency and hallucinate additional images at pseudo views. This approach improves 3D/4D scene reconstruction by explicitly e nforcing semantic correspondences during the generation of unseen views, thereby enhancing geometric consistency--unlike purely generative methods that often fa il to maintain view consistency. Extensive evaluations on various 3D/4D datasets --including Blender, LLFF, Mip-NeRF360, and Neural 3D Video--demonstrate that ou r GS-GS outperforms existing state-of-the-art methods in rendering quality witho ut sacrificing efficiency.

****

MicroVQA: A Multimodal Reasoning Benchmark for Microscopy-Based Scientific Resea rch

James Burgess, Jeffrey J Nirschl, Laura Bravo-Sánchez, Alejandro Lozano, Sanket Rajan Gupte, Jesus G. Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha Bhowmik, Zac hary Coman, Sarina M Hasan, Alexandra Johannesson, William D. Leineweber, Malvik a G Nair, Ridhi Yarlagadda, Connor Zuraski, Wah Chiu, Sarah Cohen, Jan N. Hansen , Manuel D Leonetti, Chad Liu, Emma Lundberg, Serena Yeung-Levy; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19552-1 9564

Scientific research demands sophisticated reasoning over multimodal data, a chal lenge especially prevalent in biology. Despite recent advances in multimodal lar ge language models (MLLMs) for AI-assisted research, existing multimodal reasoni ng benchmarks only target up to college-level difficulty, while research-level b enchmarks emphasize lower-level perception, falling short of the complex multimo dal reasoning needed for scientific discovery. To bridge this gap, we introduce MicroVQA, a visual-question answering (VQA) benchmark designed to assess three r easoning capabilities vital in research workflows: expert image understanding, h ypothesis generation, and experiment proposal. MicroVQA consists of 1,042 multip le-choice questions (MCQs) curated by biology experts across diverse microscopy modalities, ensuring VQA samples represent real scientific practice. In construc ting the benchmark, we find that standard MCQ generation methods induce language shortcuts, motivating a new two-stage pipeline: an optimized LLM prompt structu res question-answer pairs into MCQs; then, an agent-based 'RefineBot' updates th em to remove shortcuts. Benchmarking on state-of-the-art MLLMs reveal a peak per formance of 53%; models with smaller LLMs only slightly underperform top models, suggesting that language-based reasoning is less challenging than multimodal re asoning; and tuning with scientific articles enhances performance. Expert analys is of chain-of-thought responses shows that perception errors are the most frequ ent, followed by knowledge errors and then overgeneralization errors. These insi

ghts highlight the challenges in multimodal scientific reasoning, showing MicroVQA is a valuable resource advancing AI-driven biomedical research. MicroVQA is available at https://huggingface.co/datasets/jmhb/microvqa and project at https://jmhb0.github.io/microvqa.

****

Generative Inbetweening through Frame-wise Conditions-Driven Video Generation

Tianyi Zhu, Dongwei Ren, Qilong Wang, Xiaohe Wu, Wangmeng Zuo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27968-27978

Generative inbetweening aims to generate intermediate frame sequences by utilizing two key frames as input. Although remarkable progress has been made in video generation models, generative inbetweening still faces challenges in maintaining temporal stability due to the ambiguous interpolation path between two key frames. This issue becomes particularly severe when there is a large motion gap between input frames. In this paper, we propose a straightforward yet highly effective Frame-wise Conditions-driven Video Generation (FCVG) method that significantly enhances the temporal stability of interpolated video frames. Specifically, our FCVG provides an explicit condition for each frame, making it much easier to identify the interpolation path between two input frames and thus ensuring temporally stable production of visually plausible video frames. To achieve this, we suggest extracting matched lines from two input frames that can then be easily interpolated frame by frame, serving as frame-wise conditions seamlessly integrated into existing video generation models. In extensive evaluations covering diverse scenarios such as natural landscapes, complex human poses, camera movements and animations, existing methods often exhibit incoherent transitions across frames. In contrast, our FCVG demonstrates the capability to generate temporally stable videos using both linear and non-linear interpolation curves. Our project page and code are available at https://fcvg-inbetween.github.io/.

****

DexGrasp Anything: Towards Universal Robotic Dexterous Grasping with Physics Awareness

Yiming Zhong, Qi Jiang, Jingyi Yu, Yuexin Ma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22584-22594

A dexterous hand capable of grasping any object is essential for the development of general-purpose embodied intelligent robots. However, due to the high degree of freedom in dexterous hands and the vast diversity of objects, generating high-quality, usable grasping poses in a robust manner is a significant challenge. In this paper, we introduce DexGrasp Anything, a method that effectively integrates physical constraints into both the training and sampling phases of a diffusion-based generative model, achieving state-of-the-art performance across nearly all open datasets. Additionally, we present a new dexterous grasping dataset containing over 3.4 million diverse grasping poses for more than 15k different objects, demonstrating its potential to advance universal dexterous grasping. The code of our method and our dataset will be publicly released soon.

****

CustAny: Customizing Anything from A Single Example

Lingjie Kong, Kai Wu, Chengming Xu, Xiaobin Hu, Wenhui Han, Jinlong Peng, Donghao Luo, Mengtian Li, Jiangning Zhang, Chengjie Wang, Yanwei Fu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20916-20925

Recent advances in diffusion-based text-to-image models have simplified creating high-fidelity images, but preserving the identity (ID) of specific elements, like a personal dog, is still challenging.Object customization, using reference images and textual descriptions, is key to addressing this issue. Current object customization methods are either object-specific, requiring extensive fine-tuning, or object-agnostic, offering zero-shot customization but limited to specialized domains. The primary issue of promoting zero-shot object customization from specific domains to the general domain is to establish a large-scale general ID dataset for model pre-training, which is time-consuming and labor-intensive. In this paper, we propose a novel pipeline to construct a large dataset of general ob

jects and build the Multi-Category ID-Consistent (MC-IDC) dataset, featuring 315 k text-image samples across 10k categories. With the help of MC-IDC, we introduc e Customizing Anything (CustAny), a zero-shot framework that maintains ID fideli ty and supports flexible text editing for general objects. CustAny features thre e key components: a general ID extraction module, a dual-level ID injection modu le, and an ID-aware decoupling module, allowing it to customize any object from a single reference image and text prompt. Experiments demonstrate that CustAny o utperforms existing methods in both general object customization and specialized domains like human customization and virtual try-on. Our contributions include a large-scale dataset, the CustAny framework and novel ID processing to advance this field. The official project page is in https://lingjiekong-fdu.github.io.
****

Vision-Language Gradient Descent-driven All-in-One Deep Unfolding Networks
Haijin Zeng, Xiangming Wang, Yongyong Chen, Jingyong Su, Jie Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7524-7533
Dynamic image degradations, including noise, blur and lighting inconsistencies, pose significant challenges in image restoration, often due to sensor limitation s or adverse environmental conditions. Existing Deep Unfolding Networks (DUNs) o ffer stable restoration performance but require manual selection of degradation matrices for each degradation type, limiting their adaptability across diverse s cenarios.To address this issue, we propose the Vision-Language-guided Unfolding Network (VLU-Net), a unified DUN framework for handling multiple degradation typ es simultaneously.VLU-Net leverages a Vision-Language Model (VLM) refined on deg raded image-text pairs to align image features with degradation descriptions, se lecting the appropriate transform for target degradation.By integrating an autom atic VLM-based gradient estimation strategy into the Proximal Gradient Descent (PGD) algorithm, VLU-Net effectively tackles complex multi-degradation restoratio n tasks while maintaining interpretability. Furthermore, we design a hierarchica l feature unfolding structure to enhance VLU-Net framework, efficiently synthesi zing degradation patterns across various levels.VLU-Net is the first all-in-one DUN framework and outperforms current leading one-by-one and all-in-one end-to-e nd methods by 3.74 dB on the SOTS dehazing dataset and 1.70 dB on the Rain100L d eraining dataset.
****

3D-LLaVA: Towards Generalist 3D LMMs with Omni Superpoint Transformer
Jiajun Deng, Tianyu He, Li Jiang, Tianyu Wang, Feras Dayoub, Ian Reid; Proceedin gs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3 772-3782
Current 3D Large Multimodal Models (3D LMMs) have shown tremendous potential in 3D-vision-based dialogue and reasoning. However, how to further enhance 3D LMMs to achieve fine-grained scene understanding and facilitate flexible human-agent interaction remains a challenging problem. In this work, we introduce 3D-LLaVA, a simple yet highly powerful 3D LMM designed to act as an intelligent assistant in comprehending, reasoning, and interacting with the 3D world. Unlike existing top-performing methods that rely on complicated pipelines--such as offline multi -view feature extraction or additional task-specific heads--3D-LLaVA adopts a mi nimalist design with integrated architecture and only takes point clouds as inpu t. At the core of 3D-LLaVA is a new Omni Superpoint Transformer (OST), which int egrates three functionalities: (1) a visual feature selector that converts and s elects visual tokens, (2) a visual prompt encoder that embeds interactive visual prompts into the visual token space, and (3) a referring mask decoder that prod uces 3D masks based on text description. This versatile OST is empowered by the hybrid pretraining to obtain perception priors and leveraged as the visual conne ctor that bridges the 3D data to the LLM. After performing unified instruction t uning, our 3D-LLaVA reports impressive results on various benchmarks. The code a nd model will be released at https://github.com/djiajunustc/3D-LLaVA.
****

Event-based Video Super-Resolution via State Space Models
Zeyu Xiao, Xinchao Wang; Proceedings of the Computer Vision and Pattern Recognit

ion Conference (CVPR), 2025, pp. 12564-12574
Exploiting temporal correlations is crucial for video super-resolution (VSR). Recent approaches enhance this by incorporating event cameras. In this paper, we introduce MamEVSR, a Mamba-based network for event-based VSR that leverages the selective state space model, Mamba. MamEVSR stands out by offering global receptive field coverage with linear computational complexity, thus addressing the limitations of convolutional neural networks and Transformers. The key components of MamEVSR include: (1) The interleaved Mamba (iMamba) block, which interleaves tokens from adjacent frames and applies multidirectional selective state space modeling, enabling efficient feature fusion and propagation across bi-directional frames while maintaining linear complexity. (2) The crossmodality Mamba (cMamba) block facilitates further interaction and aggregation between event information and the output from the iMamba block. The cMamba block can leverage complementary spatio-temporal information from both modalities and allows MamEVSR to capture finermotion details. Experimental results show that the proposed MamEVSR achieves superior performance on various datasets quantitatively and qualitatively.
****

PoseTraj: Pose-Aware Trajectory Control in Video Diffusion
Longbin Ji, Lei Zhong, Pengfei Wei, Changjian Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22776-22785
Recent advancements in trajectory-guided video generation have achieved notable progress. However, existing models still face challenges in generating object motions with potentially changing 6D poses under wide-range rotations, due to limited 3D understanding. To address this problem, we introduce PoseTraj, a pose-aware video dragging model for generating 3D-aligned motion from 2D trajectories. Our method adopts a novel two-stage pose-aware pretraining framework, improving 3D understanding across diverse trajectories. Specifically, we propose a large-scale synthetic dataset PoseTraj-10k, containing 10k videos of objects following rotational trajectories, and enhance the model perception of object pose changes by incorporating 3D bounding boxes as intermediate supervision signals. Following this, we fine-tune the trajectory-controlling module on real-world videos, applying an additional camera-disentanglement module to further refine motion accuracy. Experiments on various benchmark datasets demonstrate that our method not only excels in 3D pose-aligned dragging for rotational trajectories but also outperforms existing baselines in trajectory accuracy and video quality.
****

Masked Scene Modeling: Narrowing the Gap Between Supervised and Self-Supervised Learning in 3D Scene Understanding
Pedro Hermosilla, Christian Stippel, Leon Sick; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14835-14844
Self-supervised learning has transformed 2D computer vision by enabling models trained on large, unannotated datasets to provide versatile off-the-shelf features that perform similarly to models trained with labels. However, in 3D scene understanding, self-supervised methods are typically only used as a weight initialization step for task-specific fine-tuning, limiting their utility for general-purpose feature extraction. This paper aims to address this shortcoming by proposing a robust evaluation protocol specifically designed to assess the quality of self-supervised features for 3D scene understanding. Our protocol uses multi-resolution feature sampling of hierarchical models to create rich point-level representations that capture the semantic capabilities of the model and, hence, are suitable for evaluation with linear probing and nearest-neighbor methods. Furthermore, we introduce the first self-supervised model that performs similarly to supervised models when only off-the-shelf features are used in a linear probing setup. In particular, our model is trained natively in 3D with a novel self-supervised approach based on a Masked Scene Modeling objective, which reconstructs deep features of masked patches in a bottom-up manner and is specifically tailored to hierarchical 3D models. Our experiments not only demonstrate that our method achieves competitive performance to supervised models, but also surpasses existing self-supervised approaches by a large margin.
****

VL2Lite: Task-Specific Knowledge Distillation from Large Vision-Language Models to Lightweight Networks
Jinseong Jang, Chunfei Ma, Byeongwon Lee; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30073-30083
Deploying high-performing neural networks in resource-constrained environments poses a significant challenge due to the computational demands of large-scale models. We introduce VL2Lite, a knowledge distillation framework designed to enhance the performance of lightweight neural networks in image classification tasks by leveraging the rich representational knowledge from Vision-Language Models (VLMs). VL2Lite directly integrates multi-modal knowledge from VLMs into compact models during training, effectively compensating for the limited computational and modeling capabilities of smaller networks. By transferring high-level features and complex data representations, our approach improves the accuracy and efficiency of image classification tasks without increasing computational overhead during inference. Experimental evaluations demonstrate that VL2Lite achieves up to a 7% improvement in classification performance across various datasets. This method addresses the challenge of deploying accurate models in environments with constrained computational resources, offering a balanced solution between model complexity and operational efficiency.
****
Boost the Inference with Co-training: A Depth-guided Mutual Learning Framework for Semi-supervised Medical Polyp Segmentation
Yuxin Li, Zihao Zhu, Yuxiang Zhang, Yifan Chen, Zhibin Yu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10394-10403
Semi-supervised polyp segmentation has made significant progress in recent years as a potential solution for computer-assisted treatment. Since depth images can provide extra information other than RGB images to help segment these problematic areas, depth-assisted polyp segmentation has gained much attention. However, the utilization of depth information is still worth studying. The existing RGB-D segmentation methods rely on depth data in the inference stage, limiting their clinical applications. To tackle this problem, we propose a semi-supervised polyp segmentation framework based on the mean teacher architecture. We establish an auxiliary student network with depth images as input in the training stage, and we propose a depth-guided cross-modal mutual learning strategy to promote the learning of complementary information between different student networks. Meanwhile, we use the high-confidence pseudo-labels generated by the auxiliary student network to guide the learning progress of the main student network from different perspectives. Our model does not need depth data in the inference phase. In addition, we introduce a depth-guided patch augmentation method to improve the model's learning performance in difficult regions of unlabeled polyp images. Experimental results show that our method achieves state-of-the-art performance under different label conditions on five polyp datasets. The code is available at https://github.com/pingchuan/RD-Net.
****
VidHalluc: Evaluating Temporal Hallucinations in Multimodal Large Language Models for Video Understanding
Chaoyu Li, Eun Woo Im, Pooyan Fazli; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13723-13733
Multimodal large language models (MLLMs) have recently shown significant advancements in video understanding, excelling in content reasoning and instruction-following tasks. However, hallucination, where models generate inaccurate or misleading content, remains underexplored in the video domain. Building on the observation that MLLM visual encoders often fail to distinguish visually different yet semantically similar video pairs, we introduce VidHalluc, the largest benchmark designed to examine hallucinations in MLLMs for video understanding. It consists of 5,002 videos, paired to highlight cases prone to hallucinations. VidHalluc assesses hallucinations across three critical dimensions: (1) action, (2) temporal sequence, and (3) scene transition. Comprehensive testing shows that most MLLMs are vulnerable to hallucinations across these dimensions. Furthermore, we propose DINO-HEAL, a training-free method that reduces hallucinations by incorporati

ng spatial saliency from DINOv2 to reweight visual features during inference. Our results show that DINO-HEAL consistently improves performance on VidHalluc, achieving an average improvement of 3.02% in mitigating hallucinations across all tasks. Both the VidHalluc benchmark and DINO-HEAL code are available at https://people-robots.github.io/vidhalluc.
****

StageDesigner: Artistic Stage Generation for Scenography via Theater Scripts
Zhaoxing Gan, Mengtian Li, Ruhua Chen, Zhongxia Ji, Sichen Guo, Huanling Hu, Guangnan Ye, Zuo Hu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28705-28714
In this work, we introduce StageDesigner, the first comprehensive framework for artistic stage generation using large language models combined with layout-controlled diffusion models. Given the professional requirements of stage scenography, StageDesigner simulates the workflows of seasoned artists to generate immersive 3D stage scenes. Specifically, our approach is divided into three primary modules: Script Analysis, which extracts thematic and spatial cues from input scripts; Foreground Generation, which constructs and arranges essential 3D objects; and Background Generation, which produces a harmonious background aligned with the narrative atmosphere and maintains spatial coherence by managing occlusions between foreground and background elements. Furthermore, we introduce the StagePro-V1 dataset, a dedicated dataset with 276 unique stage scenes spanning different historical styles and annotated with scripts, images, and detailed 3D layouts, specifically tailored for this task. Finally, evaluations using both standard and newly proposed metrics, along with extensive user studies, demonstrate the effectiveness of StageDesigner, showcasing its ability to produce visually and thematically cohesive stages that meet both artistic and spatial coherence standards. Project can be found at: https://deadsmither5.github.io/2025/01/03/StageDesigner/
****

From Laboratory to Real World: A New Benchmark Towards Privacy-Preserved Visible-Infrared Person Re-Identification
Yan Jiang, Hao Yu, Xu Cheng, Haoyu Chen, Zhaodong Sun, Guoying Zhao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8828-8837
Aiming to match pedestrian images captured under varying lighting conditions, visible-infrared person re-identification (VI-ReID) has drawn intensive research attention and achieved promising results. However, in real-world surveillance contexts, data is distributed across multiple devices/entities, raising privacy and ownership concerns that make existing centralized training impractical for VI-ReID. To tackle these challenges, we propose L2RW, a benchmark that brings VI-ReID closer to real-world applications. The rationale of L2RW is that integrating decentralized training into VI-ReID can address privacy concerns in scenarios with limited data-sharing regulation. Specifically, we design protocols and corresponding algorithms for different privacy sensitivity levels. In our new benchmark, we ensure the model training is done in the conditions that: 1) data from each camera remains completely isolated, or 2) different data entities (e.g., data controllers of a certain region) can selectively share the data. In this way, we simulate scenarios with strict privacy constraints which is closer to real-world conditions. Intensive experiments with various server-side federated algorithms are conducted, showing the feasibility of decentralized VI-ReID training. Notably, when evaluated in unseen domains (i.e., new data entities), our L2RW, trained with isolated data (privacy-preserved), achieves performance comparable to SOTAs trained with shared data (privacy-unrestricted). We hope this work offers a novel research entry for deploying VI-ReID that fits real-world scenarios and can benefit the community.
****

4Deform: Neural Surface Deformation for Robust Shape Interpolation
Lu Sang, Zehranaz Canfes, Dongliang Cao, Riccardo Marin, Florian Bernard, Daniel Cremers; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6542-6551

Generating realistic intermediate shapes between non-rigidly deformed shapes is a challenging task in computer vision, especially with unstructured data (e.g., point clouds) where temporal consistency across frames is lacking, and topologies are changing. Most interpolation methods are designed for structured data (i.e., meshes) and do not apply to real-world point clouds. In contrast, our approach leverages neural implicit representation (NIR) to enable free-topology changing shape deformation. Unlike previous mesh-based methods, which model learns vertex-based deformation fields, our method learns a continuous velocity field in Euclidean space, making it suitable for less structured data such as point clouds. Additionally, our method does not require intermediate-shape supervision during training; instead, we incorporate physical and geometrical constraints to regularize the velocity field. We reconstruct intermediate surfaces using a modified level-set equation, directly linking our NIR with the velocity field. Experiments show that our method significantly outperforms previous NIR approaches across various scenarios (e.g., noisy, partial, topology-changing, non-isometric shapes) and, for the first time, enables new applications like 4D Kinect sequence upsampling and real-world high-resolution mesh deformation.

****

Dense Match Summarization for Faster Two-view Estimation

Jonathan Astermark, Anders Heyden, Viktor Larsson; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1093-1102

In this paper, we speed up robust two-view relative pose from dense correspondences. Previous work has shown that dense matchers can significantly improve both accuracy and robustness in the resulting pose. However, the large number of matches comes with a significantly increased runtime during robust estimation in RANSAC. To avoid this, we propose an efficient match summarization scheme which provides comparable accuracy to using the full set of dense matches, while having 10-100x faster runtime. We validate our approach on standard benchmark datasets together with multiple state-of-the-art dense matchers.

****

Align-A-Video: Deterministic Reward Tuning of Image Diffusion Models for Consistent Video Editing

Shengzhi Wang, Yingkang Zhong, Jiangchuan Mu, Kai Wu, Mingliang Xiong, Wen Fang, Mingqing Liu, Hao Deng, Bin He, Gang Li, Qingwen Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2074-2083

Due to control limitations in the denoising process and the lack of training, zero-shot video editing methods often struggle to meet user instructions, resulting in generated videos that are visually unappealing and fail to fully satisfy expectations. To address this problem, we propose Align-A-Video, a video editing pipeline that incorporates human feedback through reward fine-tuning. Our approach consists of two key steps: 1) Deterministic Reward Fine-tuning. To reduce optimization costs for expected noise distributions, we propose a deterministic reward tuning strategy. This method improves tuning stability by increasing sample determinism, allowing the tuning process to be completed in minutes; 2) Feature Propagation Across Frames. We optimize a selected anchor frame and propagate its features to the remaining frames, improving both visual quality and semantic fidelity. This approach avoids temporal consistency degradation from reward optimization. Extensive qualitative and quantitative experiments confirm the effectiveness of using reward fine-tuning in Align-A-Video, significantly improving the overall quality of generated videos.

****

Interpreting Object-level Foundation Models via Visual Precision Search

Ruoyu Chen, Siyuan Liang, Jingzhi Li, Shiming Liu, Maosen Li, Zhen Huang, Hua Zhang, Xiaochun Cao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30042-30052

Advances in multimodal pre-training have propelled object-level foundation models, such as Grounding DINO and Florence-2, in tasks like visual grounding and object detection. However, interpreting these models' decisions has grown increasingly challenging. Existing interpretable attribution methods for object-level task interpretation have notable limitations: (1) gradient-based methods lack preci

se localization due to visual-textual fusion in foundation models, and (2) pertu rbation-based methods produce noisy saliency maps, limiting fine-grained interpr etability. To address these, we propose a Visual Precision Search method that ge nerates accurate attribution maps with fewer regions. Our method bypasses intern al model parameters to overcome attribution issues from multimodal fusion, divid ing inputs into sparse sub-regions and using consistency and collaboration score s to accurately identify critical decision-making regions. We also conducted a t heoretical analysis of the boundary guarantees and scope of applicability of our method. Experiments on RefCOCO, MS COCO, and LVIS show our approach enhances ob ject-level task interpretability over SOTA for Grounding DINO and Florence-2 acr oss various evaluation metrics, with faithfulness gains of 23.7%, 31.6%, and 20. 1% on MS COCO, LVIS, and RefCOCO for Grounding DINO, and 50.7% and 66.9% on MS C OCO and RefCOCO for Florence-2. Additionally, our method can interpret failures in visual grounding and object detection tasks, surpassing existing methods acro ss multiple evaluation metrics. The code is released at https://github.com/Ruoyu Chen10/VPS.

****

Foley-Flow: Coordinated Video-to-Audio Generation with Masked Audio-Visual Align ment and Dynamic Conditional Flows

Shentong Mo, Yibing Song; Proceedings of the Computer Vision and Pattern Recogni tion Conference (CVPR), 2025, pp. 28912-28921

Coordinated audio generation based on video inputs typically requires a strict a udio-visual (AV) alignment, where both semantics and rhythmics of the generated audio segments shall correspond to those in the video frames. Previous studies l everage a two-stage design where the AV encoders are firstly aligned via contras tive learning, then the encoded video representations guide the audio generation process. We observe that both contrastive learning and global video guidance ar e effective in aligning overall AV semantics while limiting temporally rhythmic synchronization. In this work, we propose Foley-Flow to first align unimodal AV encoders via masked modeling training, where the masked audio segments are recov ered under the guidance of the corresponding video segments. After training, the AV encoders which are separately pretrained using only unimodal data are aligne d with semantic and rhythmic consistency. Then, we develop a dynamic conditional flow for the final audio generation. Built upon the efficient velocity flow gen eration framework, our dynamic conditional flow utilizes temporally varying vide o features as the dynamic condition to guide corresponding audio segment generat ions. To this end, we extract coherent semantic and rhythmic representations dur ing masked AV alignment, and use this representation of video segments to guide audio generation temporally. Our audio results are evaluated on the standard ben chmarks and largely surpass existing results under several metrics. The superior performance indicates that Foley-Flow is effective in generating coordinated au dios that are both semantically and rhythmically coherent to various video seque nces.

****

LION-FS: Fast & Slow Video-Language Thinker as Online Video Assistant

Wei Li, Bing Hu, Rui Shao, Leyang Shen, Liqiang Nie; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3240-3251

First-person video assistants are highly anticipated to enhance our daily life t hrough online video dialogue. However, existing online video assistants often sa crifice assistant efficacy for real-time efficiency by processing low-frame-rate videos with coarse-grained visual features. To overcome the trade-off between e fficacy and efficiency, we propose "**F**ast & **S**low Video-Language Thinker" as on**LI**ne vide**O** assista**N**t, **LION-FS**, achieving real-time, proacti ve, temporally accurate, and contextually precise responses. LION-FS adopts a tw o-stage optimization strategy: **1) Fast Path: Routing-Based Response Determinat ion** evaluates frame-by-frame whether a immediate response is necessary. To enh ance responses determination accuracy and handle higher frame-rate inputs effici ently, we employ Token Aggregation Routing to dynamically fuse spatiotemporal fe atures without increasing token numbers, while utilizing Token Dropping Routing to eliminate redundant features, and **2) Slow Path: Multi-granularity Keyframe

Augmentation** optimizes keyframes during response generation. To provide comprehensive and detailed responses beyond atomic actions constrained by training data, fine-grained spatial features and human-environment interaction features are extracted through multi-granular pooling. They are further integrated into a meticulously designed multimodal Thinking Template to guide more precise response generation. Comprehensive evaluations on online video tasks demonstrate that LION-FS achieves state-of-the-art efficacy and efficiency. The codes will be released soon.

****

CARE Transformer: Mobile-Friendly Linear Visual Transformer via Decoupled Dual Interaction

Yuan Zhou, Qingshan Xu, Jiequan Cui, Junbao Zhou, Jing Zhang, Richang Hong, Hanwang Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20135-20145

Recently, large efforts have been made to design efficient linear-complexity visual Transformers. However, current linear attention models are generally unsuitable to be deployed in resource-constrained mobile devices, due to suffering from either few efficiency gains or significant accuracy drops. In this paper, we propose a new deCoupled duAl-interactive lineaR attEntion (CARE) mechanism, revealing that features' decoupling and interaction can fully unleash the power of linear attention. We first propose an asymmetrical feature decoupling strategy that asymmetrically decouples the learning process for local inductive bias and long-range dependencies, thereby preserving sufficient local and global information while effectively enhancing the efficiency of models. Then, a dynamic memory unit is employed to maintain critical information along the network pipeline. Moreover, we design a dual interaction module to effectively facilitate interaction between local inductive bias and long-range information as well as among features at different layers. By adopting a decoupled learning way and fully exploiting complementarity across features, our method can achieve both high efficiency and accuracy. Extensive experiments on ImageNet-1K, COCO, and ADE20K datasets demonstrate the effectiveness of our approach, e.g., achieving 78.4/82.1% top-1 accuracy on ImagegNet-1K at the cost of only 0.7/1.9 GMACs. Codes will be released on github.

****

Paint by Inpaint: Learning to Add Image Objects by Removing Them First

Navve Wasserman, Noam Rotstein, Roy Ganz, Ron Kimmel; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18313-18324

Image editing has advanced significantly with the introduction of text-conditioned diffusion models. Despite this progress, seamlessly adding objects to images based on textual instructions without requiring user-provided input masks remains a challenge. We address this by leveraging the insight that removing objects (Inpaint) is significantly simpler than its inverse process of adding them (Paint), attributed to inpainting models that benefit from segmentation mask guidance. Capitalizing on this realization, by implementing an automated and extensive pipeline, we curate a filtered large-scale image dataset containing pairs of images and their corresponding object-removed versions. Using these pairs, we train a diffusion model to inverse the inpainting process, effectively adding objects into images. Unlike other editing datasets, ours features natural target images instead of synthetic ones while ensuring source-target consistency by construction. Additionally, we utilize a large Vision-Language Model to provide detailed descriptions of the removed objects and a Large Language Model to convert these descriptions into diverse, natural-language instructions. Our quantitative and qualitative results show that the trained model surpasses existing models in both object addition and general editing tasks. Visit our project page for the released dataset and trained models: https://rotsteinnoam.github.io/Paint-by-Inpaint/.

****

Motion-Grounded Video Reasoning: Understanding and Perceiving Motion at Pixel Level

Andong Deng, Tongjia Chen, Shoubin Yu, Taojiannan Yang, Lincoln Spencer, Yapeng Tian, Ajmal Saeed Mian, Mohit Bansal, Chen Chen; Proceedings of the Computer Vis

ion and Pattern Recognition Conference (CVPR), 2025, pp. 8625-8636
In this paper, we introduce Motion-Grounded Video Reasoning, a new motionunderstanding task that requires generating visual answers (video segmentationmasks) according to the input question, and hence needs implicit spatiotemporalreasoning and grounding. This task extends existing spatiotemporal groundingwork focusing on explicit action/motion grounding, to a more general format byenabling implicit reasoning via questions. To facilitate the development of the newtask, we collect a large-scale dataset called GROUNDMORE, which comprises1,715 video clips, 249K object masks that are deliberately designed with 4 questiontypes (Causal, Sequential, Counterfactual, and Descriptive) for benchmarkingdeep and comprehensive motion reasoning abilities. GROUNDMORE uniquelyrequires models to generate visual answers, providing a more concrete and visuallyinterpretable response than plain texts. It evaluates models on both spatiotemporalgrounding and reasoning, fostering to address complex challenges in motion-relatedvideo reasoning, temporal perception, and pixel-level understanding. Furthermore,we introduce a novel baseline model named Motion-Grounded Video ReasoningAssistant (MORA). MORA incorporates the multimodal reasoning ability from theMultimodal LLM, the pixel-level perception capability from the grounding model(SAM), and the temporal perception ability from a lightweight localization head.MORA achieves respectable performance on GROUNDMORE outperforming thebest existing visual grounding baseline model by an average of 21.5% relatively.We hope this novel and challenging task will pave the way for future advancementsin robust and general motion understanding via video reasoning segmentation.
****

PMA: Towards Parameter-Efficient Point Cloud Understanding via Point Mamba Adapter
Yaohua Zha, Yanzi Wang, Hang Guo, Jinpeng Wang, Tao Dai, Bin Chen, Zhihao Ouyang, Xue Yuerong, Ke Chen, Shu-Tao Xia; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16976-16986
Applying pre-trained models to assist point cloud understanding has recently become a mainstream paradigm in 3D perception. However, existing application strategies are straightforward, utilizing only the final output of the pre-trained model for various task heads. It neglects the rich complementary information in the intermediate layer, thereby failing to fully unlock the potential of pre-trained models. To overcome this limitation, we propose an orthogonal solution: Point Mamba Adapter (PMA), which constructs an ordered feature sequence from all layers of the pre-trained model and leverages Mamba to fuse all complementary semantics, thereby promoting comprehensive point cloud understanding. Constructing this ordered sequence is non-trivial due to the inherent isotropy of 3D space. Therefore, we further propose a geometry-constrained gate prompt generator (G2PG) shared across different layers, which applies shared geometric constraints to the output gates of the Mamba and dynamically optimizes the spatial order, thus enabling more effective integration of multi-layer information. Extensive experiments conducted on challenging point cloud datasets across various tasks demonstrate that our PMA elevates the capability for point cloud understanding to a new level by fusing diverse complementary intermediate features. Code is available at https://github.com/zyh16143998882/PMA.
****

All-directional Disparity Estimation for Real-world QPD Images
Hongtao Yu, Shaohui Song, Lihu Sun, Wenkai Su, Xiaodong Yang, Chengming Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21836-21846
Quad Photodiode (QPD) sensors represent an evolution by providing four sub-views, whereas dual-pixel (DP) sensors are limited to two sub-views. In addition to enhancing auto-focus performance, QPD sensors also enable disparity estimation in horizontal and vertical directions. However, the characteristics of QPD sensors, including uneven illumination across sub-views and the narrow baseline, render algorithm design difficult. Furthermore, effectively utilizing the two-directional disparity of QPD sensors remains a challenge. The scarcity of QPD disparity datasets also limits the development of learning-based methods. In this work, we

address these challenges by first proposing a DPNet for DP disparity estimation. Specifically, we design an illumination-invariant module to reduce the impact of illumination, followed by a coarse-to-fine module to estimate sub-pixel disparity. Building upon the DPNet, we further propose a QuadNet, which integrates the two-directional disparity via an edge-aware fusion module. To facilitate the evaluation of our approaches, we propose the first QPD disparity dataset QPD2K, comprising 2,100 real-world QPD images and corresponding disparity maps. Experiments demonstrate that our approaches achieve state-of-the-art performance in DP and QPD disparity estimation.

****

LC-Mamba: Local and Continuous Mamba with Shifted Windows for Frame Interpolation

Min Wu Jeong, Chae Eun Rhee; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17671-17681

In this paper, we propose LC-Mamba, a Mamba-based model that captures fine-grained spatiotemporal information in video frames, addressing limitations in current interpolation methods and enhancing performance. The main contributions are as follows: First, we apply a shifted local window technique to reduce historical decay and enhance local spatial features, allowing multiscale capture of detailed motion between frames. Second, we introduce a Hilbert curve-based selective state scan to maintain continuity across window boundaries, preserving spatial correlations both within and between windows. Third, we extend the Hilbert curve to enable voxel-level scanning to effectively capture spatiotemporal characteristics between frames. The proposed LC-Mamba achieves competitive results, with a PSNR of 36.53 dB on Vimeo-90k, outperforming prior models by +0.03 dB. The code and models are publicly available at https://github.com/Miinuuu/LC-Mamba.git

****

Zero-Shot Head Swapping in Real-World Scenarios

Taewoong Kang, Sohyun Jeong, Hyojin Jang, Jaegul Choo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10805-10814

With growing demand in media and social networks for personalized images, the need for advanced head-swapping techniques--integrating an entire head from the head image with the body from the body image--has increased. However, traditional head-swapping methods heavily rely on face-centered cropped data with primarily frontal-facing views, which limits their effectiveness in real-world applications. Additionally, their masking methods, designed to indicate regions requiring editing, are optimized for these types of dataset but struggle to achieve seamless blending in complex situations, such as when the original data includes features like long hair extending beyond the masked area. To overcome these limitations and enhance adaptability in diverse and complex scenarios, we propose a novel head swapping method, HID, that is robust to images including the full head and the upper body, and handles from frontal to side views, while automatically generating context-aware masks. For automatic mask generation, we introduce the IOMask, which enables seamless blending of the head and body, effectively addressing integration challenges. We further introduce the hair injection module to capture hair details with greater precision. Our experiments demonstrate that the proposed approach achieves state-of-the-art performance in head swapping, providing visually consistent and realistic results across a wide range of challenging conditions.

****

Toward Robust Neural Reconstruction from Sparse Point Sets

Amine Ouasfi, Shubhendu Jena, Eric Marchand, Adnane Boukhayma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6552-6562

We consider the challenging problem of learning Signed Distance Functions (SDF) from sparse and noisy 3D point clouds. In contrast to recent methods that depend on smoothness priors, our method, rooted in a distributionally robust optimization (DRO) framework, incorporates a regularization term that leverages samples from the uncertainty regions of the model to improve the learned SDFs. Thanks to tractable dual formulations, we show that this framework enables a stable and efficient optimization of SDFs in the absence of ground truth supervision. Using

a variety of synthetic and real data evaluations from different modalities, we s
how that of our DRO based learning framework can improve SDF learning with res
pect to baselines and the state-of-the-art.
****

GPAvatar: High-fidelity Head Avatars by Learning Efficient Gaussian Projections
Wei-Qi Feng, Dong Han, Ze-Kang Zhou, Shunkai Li, Xiaoqiang Liu, Pengfei Wan, Di
Zhang, Miao Wang; Proceedings of the Computer Vision and Pattern Recognition Con
ference (CVPR), 2025, pp. 250-259
Existing radiance field-based head avatar methods have mostly relied on pre-comp
uted explicit priors (e.g., mesh, point) or neural implicit representations, mak
ing it challenging to achieve high fidelity with both computational efficiency a
nd low memory consumption. To overcome this, we present GPAvatar, a novel and ef
ficient Gaussian splatting-based method for reconstructing high-fidelity dynamic
 3D head avatars from monocular videos. We extend Gaussians in 3D space to a hig
h-dimensional embedding space encompassing Gaussian's spatial position and avata
r expression, enabling the representation of the head avatar with arbitrary pose
 and expression. To enable splatting-based rasterization, a linear transformatio
n is learned to project each high-dimensional Gaussian back to the 3D space, whi
ch is sufficient to capture expression variations instead of using complex neura
l networks. Furthermore, we propose an adaptive densification strategy that dyna
mically allocates Gaussians to regions with high expression variance, improving
the facial detail representation. Experimental results on three datasets show t
hat our method outperforms existing state-of-the-art methods in rendering qualit
y and speed while reducing memory usage in training and rendering.
****

PIAD: Pose and Illumination agnostic Anomaly Detection
Kaichen Yang, Junjie Cao, Zeyu Bai, Zhixun Su, Andrea Tagliasacchi; Proceedings
of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4734
-4743
We introduce the Pose and Illumination agnostic Anomaly Detection (PIAD) problem
, a generalization of pose-agnostic anomaly detection (PAD). Being illumination
agnostic is critical, as it relaxes the assumption that training data for an obj
ect has to be acquired in the same light configuration of the query images that
we want to test. Moreover, even if the object is placed within the same capture
environment, being illumination agnostic implies that we can relax the assumptio
n that the relative pose between environment light and query object has to match
 the one in the training data. We introduce a new dataset to study this problem,
 containing both synthetic and real-world examples, propose a new baseline for P
IAD, and demonstrate how our baseline provides state-of-the-art results in both
PAD and PIAD, not only in the new proposed dataset, but also in existing dataset
s that were designed for the simpler PAD problem. Project page: https://kaichen-
yang.github.io/piad/.
****

CAV-MAE Sync: Improving Contrastive Audio-Visual Mask Autoencoders via Fine-Grai
ned Alignment
Edson Araujo, Andrew Rouditchenko, Yuan Gong, Saurabhchand Bhati, Samuel Thomas,
 Brian Kingsbury, Leonid Karlinsky, Rogerio Feris, James R. Glass, Hilde Kuehne;
 Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2
025, pp. 18794-18803
Recent advances in audio-visual learning have shown promising results in learnin
g representations across modalities. However, most approaches rely on global au
dio representations that fail to capture fine-grained temporal correspondences w
ith visual frames.Additionally, existing methods often struggle with conflicting
 optimization objectives when trying to jointly learn reconstruction and cross-m
odal alignment. In this work, we propose CAV-MAE Sync as a simple yet effective
extension of the original CAV-MAE framework for self-supervised audio-visual lea
rning. We address three key challenges: First, we tackle the granularity mismatc
h between modalities by treating audio as a temporal sequence aligned with video
 frames, rather than using global representations. Second, we resolve conflictin
g optimization goals by separating contrastive and reconstruction objectives thr

ough dedicated global tokens. Third, we improve spatial localization by introduc ing learnable register tokens that reduce semantic load on patch tokens. We eval uate the proposed approach on AudioSet, VGG Sound, and the ADE20K Sound dataset on zero-shot retrieval, classification and localization tasks demonstrating stat e-of-the-art performance and outperforming more complex architectures. Code avai lable at https://github.com/edsonroteia/cav-mae-sync.
****

Two is Better than One: Efficient Ensemble Defense for Robust and Compact Model s

Yoojin Jung, Byung Cheol Song; Proceedings of the Computer Vision and Pattern Re cognition Conference (CVPR), 2025, pp. 9696-9706

Deep learning-based computer vision systems adopt complex and large architecture s to improve performance, yet they face challenges in deployment on resource-con strained mobile and edge devices. To address this issue, model compression techn iques such as pruning, quantization, and matrix factorization have been proposed ; however, these compressed models are often highly vulnerable to adversarial at tacks. We introduce the Efficient Ensemble Defense (EED) technique, which divers ifies the compression of a single base model based on different pruning importan ce scores and enhances ensemble diversity to achieve high adversarial robustness and resource efficiency. EED dynamically determines the number of necessary sub -models during the inference stage, minimizing unnecessary computations while ma intaining high robustness. On the CIFAR-10 and SVHN datasets, EED demonstrated s tate-of-the-art robustness performance compared to existing adversarial pruning techniques, along with an inference speed improvement of up to 1.86 times. This proves that EED is a powerful defense solution in resource-constrained environme nts.
****

Tiled Diffusion

Or Madar, Ohad Fried; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7795-7804

Image tiling--the seamless connection of disparate images to create a coherent v isual field--is crucial for applications such as texture creation, video game as set development, and digital art. Traditionally, tiles have been constructed man ually, a method that poses significant limitations in scalability and flexibilit y. Recent research has attempted to automate this process using generative model s. However, current approaches primarily focus on tiling textures and manipulati ng models for single-image generation, without inherently supporting the creatio n of multiple interconnected tiles across diverse domains.This paper presents Ti led Diffusion, a novel approach that extends the capabilities of diffusion model s to accommodate the generation of cohesive tiling patterns across various domai ns of image synthesis that require tiling. Our method supports a wide range of t iling scenarios, from self-tiling to complex many-to-many connections, enabling seamless integration of multiple images.Tiled Diffusion automates the tiling pro cess, eliminating the need for manual intervention and enhancing creative possib ilities in various applications, such as seamlessly tiling of existing images, t iled texture creation, and 360deg synthesis.
****

Using Diffusion Priors for Video Amodal Segmentation

Kaihua Chen, Deva Ramanan, Tarasha Khurana; Proceedings of the Computer Vision a nd Pattern Recognition Conference (CVPR), 2025, pp. 22890-22900

Object permanence in humans is a fundamental cue that helps in understanding per sistence of objects, even when they are fully occluded in the scene. Present day methods in object segmentation do not account for this amodal nature of the wor ld, and only work for segmentation of visible or modal objects. Few amodal metho ds exist; single-image segmentation methods cannot handle high-levels of occlusi ons which are better inferred using temporal information, and multi-frame method s have focused solely on segmenting rigid objects. To this end, we propose to ta ckle video amodal segmentation by formulating it as a conditional generation tas k, thereby capitalizing on the foundational knowledge in video generative models . Our method is simple; we repurpose these models to condition on a sequence of

modal mask frames of an object along with contextual depth maps, to learn which object boundary may be occluded and therefore, extended to hallucinate the complete extent of an object. This is followed by a content completion stage which is able to inpaint the occluded regions of an object. We benchmark our approach alongside a wide array of state-of-the-art methods on four datasets and show a dramatic improvement of upto 13% for amodal segmentation in an object's occluded region.
****
COBRA: COmBinatorial Retrieval Augmentation for Few-Shot Adaptation
Arnav M. Das, Gantavya Bhatt, Lilly Kumari, Sahil Verma, Jeff Bilmes; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20534-20546
Retrieval augmentation, the practice of retrieving additional data from large auxiliary pools, has emerged as an effective technique for enhancing model performance in the low-data regime. Prior approaches have employed only nearest-neighbor based strategies for data selection, which retrieve auxiliary samples with high similarity to instances in the target task. However, these approaches are prone to selecting highly redundant samples, since they fail to incorporate any notion of diversity. In our work, we first demonstrate that data selection strategies used in prior retrieval-augmented few-shot adaptation settings can be generalized using a class of functions known as Combinatorial Mutual Information (CMI) measures. We then propose COBRA (COmBinatorial Retrieval Augmentation), which employs an alternative CMI measure that considers both diversity and similarity to a target dataset. COBRA consistently outperforms previous retrieval approaches across image classification tasks and few-shot learning techniques when used to retrieve samples from LAION-2B. COBRA introduces negligible computational overhead to the cost of retrieval while providing significant gains in downstream model performance.
****
Dyn-HaMR: Recovering 4D Interacting Hand Motion from a Dynamic Camera
Zhengdi Yu, Stefanos Zafeiriou, Tolga Birdal; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27716-27726
We propose Dyn-HaMR, to the best of our knowledge, the first approach to reconstruct 4D global hand motion from monocular videos recorded by dynamic cameras in the wild. Reconstructing accurate 3D hand meshes from monocular videos is a crucial task for understanding human behaviour, with significant applications in augmented and virtual reality (AR/VR). However, existing methods for monocular hand reconstruction typically rely on a weak perspective camera model, which simulates hand motion within a limited camera frustum. As a result, these approaches struggle to recover the full 3D global trajectory and often produce noisy or incorrect depth estimations, particularly when the video is captured by dynamic or moving cameras, which is common in egocentric scenarios. Our \name consists of a multi-stage, multi-objective optimization pipeline, that factors in (i) simultaneous localization and mapping (SLAM) to robustly estimate relative camera motion, (ii) an interacting-hand prior for generative infilling and to refine the interaction dynamics, ensuring plausible recovery under (self-)occlusions, and (iii) hierarchical initialization through a combination of state-of-the-art hand tracking methods.
****
The Scene Language: Representing Scenes with Programs, Words, and Embeddings
Yunzhi Zhang, Zizhang Li, Matt Zhou, Shangzhe Wu, Jiajun Wu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24625-24634
We introduce the Scene Language, a visual scene representation that concisely and precisely describes the structure, semantics, and identity of visual scenes. It represents a scene with three key components: a program that specifies the hierarchical and relational structure of entities in the scene, words in natural language that summarize the semantic class of each entity, and embeddings that capture the visual identity of each entity. This representation can be inferred from pre-trained language models via a training-free inference technique, given text or image inputs. The resulting scene can be rendered into images using traditi

onal, neural, or hybrid graphics renderers. Together, this forms an automated sy
stem for high-quality 3D and 4D scene generation. Compared with existing represe
ntations like scene graphs, our proposed Scene Language generates complex scenes
 with higher fidelity, while explicitly modeling the scene structures to enable
precise control and editing.
****

ProbeSDF: Light Field Probes For Neural Surface Reconstruction
Briac Toussaint, Diego Thomas, Jean-Sébastien Franco; Proceedings of the Compute
r Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11026-11035
SDF-based differential rendering frameworks have achieved state-of-the-art multi
view 3D shape reconstruction. In this work, we re-examine this family of approac
hes by minimally reformulating its core appearance model in a way that simultane
ously yields faster computation and increased performance. To this goal, we exhi
bit a physically-inspired minimal radiance parametrization decoupling angular an
d spatial contributions, by encoding them with a small number of features stored
 in two respective volumetric grids of different resolutions. Requiring as littl
e as four parameters per voxel, and a tiny MLP call inside a single fully fused
kernel, our approach allows to enhance performance with both surface and image (
PSNR) metrics, while providing a significant training speedup and real-time rend
ering. We show this performance to be consistently achieved on real data over tw
o widely different and popular application fields, generic object and human subj
ect shape reconstruction, using four representative and challenging datasets.
****

Descriptor-In-Pixel : Point-Feature Tracking For Pixel Processor Arrays
Laurie Bose, Jianing Chen, Piotr Dudek; Proceedings of the Computer Vision and P
attern Recognition Conference (CVPR), 2025, pp. 5392-5400
This paper presents a novel approach for joint point-feature detection and track
ing, designed specifically for Pixel Processor Array (PPA) vision sensors. Inste
ad of standard pixels, PPA sensors consist of thousands of "pixel-processors", e
nabling massive parallel computation of visual data at the point of light captur
e. Our approach performs all computation entirely in-pixel, meaning no raw image
 data need ever leave the sensor for external processing. We introduce a Descrip
tor-In-Pixel paradigm, in which a feature descriptor is held within the memory o
f each pixel-processor. The PPA's architecture enables the response of every pro
cessor's descriptor, upon the current image, to be computed in parallel. This pr
oduces a "descriptor response map", which, by generating the correct layout of d
escriptors across the pixel-processors, can be used for both point-feature detec
tion and tracking. This reduces sensor output to just sparse feature locations a
nd descriptors, read-out via an address-event interface, giving a greater than 1
000X reduction in data transfer compared to raw image output. The sparse readout
 and complete utilization of all pixel-processors makes our approach very effici
ent. Our implementation upon the SCAMP-7 PPA prototype runs at over 3000 FPS (Fr
ames Per Second), tracking point-features reliably under violent motion. This is
 the first work performing point-feature detection and tracking entirely in-pixe
l.
****

Hybrid Concept Bottleneck Models
Yang Liu, Tianwei Zhang, Shi Gu; Proceedings of the Computer Vision and Pattern
Recognition Conference (CVPR), 2025, pp. 20179-20189
Concept Bottleneck Models (CBMs) provide an interpretable framework for neural n
etworks by mapping visual features to predefined, human-understandable concepts.
 However, the application of CBMs is often constrained by insufficient concept a
nnotations. Recently, multi-modal pre-trained models have shown promise in reduc
ing annotation costs by aligning visual representations with textual concept emb
eddings. Nevertheless, the quality and completeness of the predefined concepts s
ignificantly affect the performance of CBMs.In this work, we propose Hybrid Conc
ept Bottleneck Model (HybridCBM), a novel CBM framework to address the challenge
 of incomplete predefined concepts. Our method consists of two main components:
a Static Concept Bank and a Dynamic Concept Bank. The Static Concept Bank direct
ly leverages large language models (LLMs) for concept construction, while the Dy

namic Concept Bank employs learnable vectors to capture complementary and valuable concepts continuously during training. After training, a pre-trained translator converts these vectors into human-understandable concepts, further enhancing model interpretability. Notably, HybridCBM is highly flexible and can be easily applied to any CBM to improve performance. Experimental results across multiple datasets demonstrate that HybridCBM outperforms current state-of-the-art methods and achieves comparable results to black-box models. Additionally, we propose novel metrics to evaluate the quality of the learned concepts, showing that they perform comparably to predefined concepts.
****

UVGS: Reimagining Unstructured 3D Gaussian Splatting using UV Mapping
Aashish Rai, Dilin Wang, Mihir Jain, Nikolaos Sarafianos, Kefan Chen, Srinath Sridhar, Aayush Prakash; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5927-5937
3D Gaussian Splatting (3DGS) has demonstrated superior quality in modeling 3D objects and scenes. However, generating 3DGS remains challenging due to their discrete, unstructured, and permutation-invariant nature. In this work, we present a simple yet effective method to overcome these challenges. We utilize spherical mapping to transform 3DGS into a structured 2D representation, termed UVGS. UVGS can be viewed as multi-channel images, with feature dimensions as a concatenation of Gaussian attributes such as position, scale, color, opacity, and rotation. We further find that these heterogeneous features can be compressed into a lower-dimensional (e.g., 3-channel) shared feature space using a carefully designed multi-branch network.The compressed UVGS can be treated as typical RGB images. Remarkably, we discover that typical VAEs trained with latent diffusion models can directly generalize to this new representation without additional training. Our novel representation makes it effortless to leverage foundational 2D models, such as diffusion models, to directly model 3DGS. Additionally, one can simply increase the 2D UV resolution to accommodate more Gaussians, making UVGS a scalable solution compared to typical 3D backbones. This approach immediately unlocks various novel generation applications of 3DGS by inherently utilizing the already developed superior 2D generation capabilities. In our experiments, we demonstrate various unconditional, conditional generation, and inpainting applications of 3DGS based on diffusion models, which were previously non-trivial.
****

Dual Consolidation for Pre-Trained Model-Based Domain-Incremental Learning
Da-Wei Zhou, Zi-Wen Cai, Han-Jia Ye, Lijun Zhang, De-Chuan Zhan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20547-20557
Domain-Incremental Learning (DIL) involves the progressive adaptation of a model to new concepts across different domains. While recent advances in pre-trained models provide a solid foundation for DIL, learning new concepts often results in the catastrophic forgetting of pre-trained knowledge. Specifically, sequential model updates can overwrite both the representation and the classifier with knowledge from the latest domain. Thus, it is crucial to develop a representation and corresponding classifier that accommodate all seen domains throughout the learning process. To this end, we propose DUal ConsolidaTion (Duct) to unify and consolidate historical knowledge at both the representation and classifier levels. By merging the backbone of different stages, we create a representation space suitable for multiple domains incrementally. The merged representation serves as a balanced intermediary that captures task-specific features from all seen domains. Additionally, to address the mismatch between consolidated embeddings and the classifier, we introduce an extra classifier consolidation process. Leveraging class-wise semantic information, we estimate the classifier weights of old domains within the latest embedding space. By merging historical and estimated classifiers, we align them with the consolidated embedding space, facilitating incremental classification. Extensive experimental results on four benchmark datasets demonstrate Duct's state-of-the-art performance. Code is available at: https://github.com/Estrella-fugaz/CVPR25-Duct.
****

Learning Physics-Based Full-Body Human Reaching and Grasping from Brief Walking References
Yitang Li, Mingxian Lin, Zhuo Lin, Yipeng Deng, Yue Cao, Li Yi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27673-27682
Existing motion generation methods based on mocap data are often limited by data quality and coverage. In this work, we propose a framework that generates diverse, physically feasible full-body human reaching and grasping motions using only brief walking mocap data. Base on the observation that walking data captures valuable movement patterns transferable across tasks and, on the other hand, the advanced kinematic methods can generate diverse grasping poses, which can then be interpolated into motions to serve as task-specific guidance. Our approach incorporates an active data generation strategy to maximize the utility of the generated motions, along with a local feature alignment mechanism that transfers natural movement patterns from walking data to enhance both the success rate and naturalness of the synthesized motions. By combining the fidelity and stability of natural walking with the flexibility and generalizability of task-specific generated data, our method demonstrates strong performance and robust adaptability in diverse scenes and with unseen objects.
****
EmoEdit: Evoking Emotions through Image Manipulation
Jingyuan Yang, Jiawei Feng, Weibin Luo, Dani Lischinski, Daniel Cohen-Or, Hui Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24690-24699
Affective Image Manipulation (AIM) seeks to modify user-provided images to evoke specific emotions. This task is inherently complex due to its twofold objective: evoking the intended emotion while preserving image composition. Existing AIM methods primarily adjust color and style, often failing to elicit precise, profound emotional shifts. Drawing on psychological insights, we introduce EmoEdit, which extends AIM by incorporating content modifications to enhance emotional impact. Specifically, we construct EmoEditSet, a large-scale AIM dataset of 40,120 paired data through emotion attribution and data construction. To make generative models emotion-aware, we design an Emotion Adapter and train it using EmoEditSet. We further propose an instruction loss to capture semantic variations in each data pair. Our method is evaluated both qualitatively and quantitatively, demonstrating superior performance over state-of-the-art techniques. Additionally, we showcase the portability of our Emotion Adapter to other diffusion-based models, enhancing their emotion knowledge with diverse semantics. Code is available at: https://github.com/JingyuanYY/EmoEdit.
****
RORem: Training a Robust Object Remover with Human-in-the-Loop
Ruibin Li, Tao Yang, Song Guo, Lei Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14024-14035
Despite the significant advancements, existing object removal methods struggle with incomplete removal, incorrect content synthesis and blurry synthesized regions, resulting in low success rates. Such issues are mainly caused by the lack of high-quality paired training data, as well as the self-supervised training paradigm adopted in these methods, which forces the model to in-paint the masked regions, leading to ambiguity between synthesizing the masked objects and restoring the background. To address these issues, we propose a semi-supervised learning strategy with human-in-the-loop to create high-quality paired training data, aiming to train a Robust Object Remover (RORem). We first collect 60K training pairs from open-source datasets to train an initial object removal model for generating removal samples, and then utilize human feedback to select a set of high-quality object removal pairs, with which we train a discriminator to automate the following training data generation process. By iterating this process for several rounds, we finally obtain a substantial object removal dataset with over 200K pairs. Fine-tuning the pre-trained stable diffusion model with this dataset, we obtain our RORem, which demonstrates state-of-the-art object removal performance in terms of both reliability and image quality. Particularly, RORem improves the

object removal success rate over previous methods by more than 18%. The dataset, source code and trained model will be released.

****

All Languages Matter: Evaluating LMMs on Culturally Diverse 100 Languages

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kukreja, Mykola Maslych, Wafa Al Ghallabi, Mihail Minkov Mihaylov, Chao Qin, Abdelrahman M. Shaker, Mike Zhang, Mahardika Krisna Ihsani, Amiel Gian Esplana, Monil Gokani, Shachar Mirkin, Harsh Singh, Ashay Srivastava, Endre Hamerlik, Fathinah Asma Izzati, Fadillah Adamsyah Maani, Sebastian Cavada, Jenny Chim, Rohit Gupta, Sanjay Manjunath, Kamila Zhumakhanova, Feno Heriniaina Rabevohitra, Azril Hafizi Amirudin, Muhammad Ridzuan, Daniya Najiha Abdul Kareem, Ketan Pravin More, Kunyang Li, Pramesh Shakya, Muhammad Saad, Amirpouya Ghasemaghaei, Amirbek Djanibekov, Dilshod Azizov, Branislava Jankovic, Naman Bhatia, Alvaro Cabrera, Johan Obando-Ceron, Olympiah Otieno, Febian Farestam, Muztoba Rabbani, Sanoojan Ballah, Santosh Sanjeev, Abduragim Shtanchaev, Maheen Fatima, Thao Nguyen, Amrin Kareem, Toluwani Aremu, Nathan Augusto Zacarias Xavier, Amit Bhatkal, Hawau Olamide Toyin, Aman Chadha, Hisham Cholakkal, Rao Muhammad Anwer, Michael Felsberg, Jorma Laaksonen, Thamar Solorio, Monojit Choudhury, Ivan Laptev, Mubarak Shah, Salman Khan, Fahad Shahbaz Khan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19565-19575

Existing Large Multimodal Models (LMMs) generally focus on only a few regions and languages. As LMMs continue to improve, it is increasingly important to ensure they understand cultural contexts, respect local sensitivities, and support low-resource languages, all while effectively integrating corresponding visual cues. In pursuit of culturally diverse global multimodal models, our proposed All Languages Matter Benchmark (ALM-bench) represents the largest and most comprehensive effort to date for evaluating LMMs across 100 languages. ALM-bench challenges existing models by testing their ability to understand and reason about culturally diverse images paired with text in various languages, including many low-resource languages traditionally underrepresented in multimodal research. The benchmark offers a robust and nuanced evaluation framework featuring various question formats, including True/False, multiple choice, and open-ended questions, which are further divided into short and long-answer categories. ALM-bench design ensures a comprehensive assessment of a model's ability to handle varied levels of difficulty in visual and linguistic reasoning. To capture the rich tapestry of global cultures, ALM-bench carefully curates content from 13 distinct cultural aspects, ranging from traditions and rituals to famous personalities and celebrations. Through this, ALM-bench not only provides a rigorous testing ground for state-of-the-art open and closed-source LMMs but also highlights the importance of cultural and linguistic inclusivity, encouraging the development of models that can serve diverse global populations effectively. Our benchmark will be publicly released.

****

SparseAlign: a Fully Sparse Framework for Cooperative Object Detection

Yunshuang Yuan, Yan Xia, Daniel Cremers, Monika Sester; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22296-22305

Cooperative perception can increase the view field and decrease the occlusion of an ego vehicle, hence improving the perception performance and safety of autonomous driving. Despite the success of previous works on cooperative object detection, they mostly operate on dense Bird's Eye View (BEV) feature maps, which are computationally demanding and can hardly be extended to long-range detection problems. More efficient fully sparse frameworks are rarely explored. In this work, we design a fully sparse framework, SparseAlign, with three key features: an enhanced sparse 3D backbone, a query-based temporal context learning module, and a robust detection head specially tailored for sparse features. Extensive experimental results on both OPV2V and DairV2X datasets show that our framework, despite its sparsity, outperforms the state of the art with less communication bandwidth requirements. In addition, experiments on the OPV2Vt and DairV2Xt datasets for time-aligned cooperative object detection also show a significant performance

gain compared to the baseline works.
****
Video-Bench: Human-Aligned Video Generation Benchmark
Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Yufan Deng, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, Jie Zhang, Chi Zhang, Li-jia Li, Yongxin Ni; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18858-18868
Video generation assessment is essential for ensuring that generative models produce visually realistic, high-quality videos while aligning with human expectations. Current video generation benchmarks fall into two main categories: traditional benchmarks, which use metrics and embeddings to evaluate generated video quality across multiple dimensions but often lack alignment with human judgments; and large language model (LLM)-based benchmarks, though capable of human-like reasoning, are constrained by a limited understanding of video quality metrics and cross-modal consistency.To address these challenges and establish a benchmark that better aligns with human preferences, this paper introduces Video-Bench, a comprehensive benchmark featuring a rich prompt suite and extensive evaluation dimensions. This benchmark represents the first attempt to systematically leverage MLLMs across all dimensions relevant to video generation assessment in generative models. By incorporating few-shot scoring and chain-of-query techniques, Video-Bench provides a structured, scalable approach to generated video evaluation. Experimental results demonstrate that MLLMs achieve superior alignment with human preferences across all dimensions. Moreover, in instances where our framework's assessments diverge from human evaluations, it consistently offers more objective and accurate insights, suggesting an even greater potential advantage over traditional human judgment.
****
Data Distributional Properties As Inductive Bias for Systematic Generalization
Felipe del Rio, Alain Raymond-Saez, Daniel Florea, Rodrigo Toro Icarte, Julio Hurtado, Cristian B. Calderon, Alvaro Soto; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25590-25601
Deep neural networks (DNNs) struggle at systematic generalization (SG). Several studies have evaluated the possibility of promoting SG through the proposal of novel architectures, loss functions, or training methodologies. Few studies, however, have focused on the role of training data properties in promoting SG. In this work, we investigate the impact of certain data distributional properties, as inductive biases for the SG ability of a multi-modal language model. To this end, we study three different properties. First, data diversity, instantiated as an increase in the possible values a latent property in the training distribution may take. Second, burstiness, where we probabilistically restrict the number of possible values of latent factors on particular inputs during training. Third, latent intervention, where a particular latent factor is altered randomly during training. We find that all three factors significantly enhance SG, with diversity contributing an 89% absolute increase in accuracy in the most affected property. Through a series of experiments, we test various hypotheses to understand why these properties promote SG. Finally, we find that Normalized Mutual Information (NMI) between latent attributes in the training distribution is strongly predictive of out-of-distribution generalization. We find that a mechanism by which lower NMI induces SG is in the geometry of representations. In particular, we find that NMI induces more parallelism in neural representations (i.e., input features coded in parallel neural vectors) of the model, a property related to the capacity of reasoning by analogy.
****
MergeVQ: A Unified Framework for Visual Generation and Representation with Disentangled Token Merging and Quantization
Siyuan Li, Luyuan Zhang, Zedong Wang, Juanxi Tian, Cheng Tan, Zicheng Liu, Chang Yu, Qingsong Xie, Haonan Lu, Haoqian Wang, Zhen Lei; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19713-19723
Masked Image Modeling (MIM) with Vector Quantization (VQ) has achieved great success in both self-supervised pre-training and image generation. However, most ex

isting methods struggle to address the trade-off in the shared latent space for generation quality vs. representation learning and efficiency. To push the limits of this paradigm, we propose MergeVQ, which incorporates token merging techniques into VQ-based generative models to bridge the gap between image generation and visual representation learning in a unified architecture. During pre-training, MergeVQ decouples top-k semantics from latent space with the token merge module after self-attention blocks in the encoder for subsequent Look-up Free Quantization (LFQ) and global alignment and recovers their fine-grained details through cross-attention in the decoder for reconstruction. As for second-stage generation, we introduce MergeAR, which performs KV Cache compression for efficient raster-order prediction. Extensive experiments on ImageNet verify that MergeVQ as an AR generative model achieves competitive performance in both visual representation learning and image generation tasks while maintaining favorable token efficiency and inference speed. The code and model will be available at https://apexgen-x.github.io/MergeVQ.

****

InterAct: Advancing Large-Scale Versatile 3D Human-Object Interaction Generation
Sirui Xu, Dongting Li, Yucheng Zhang, Xiyan Xu, Qi Long, Ziyin Wang, Yunzhi Lu, Shuchang Dong, Hezi Jiang, Akshat Gupta, Yu-Xiong Wang, Liang-Yan Gui; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7048-7060

While large-scale human motion capture datasets have advanced human motion generation, modeling and generating dynamic 3D human-object interactions (HOIs) remain challenging due to dataset limitations. Existing datasets often lack extensive, high-quality motion and annotation and exhibit artifacts such as contact penetration, floating, and incorrect hand motions. To address these issues, we introduce InterAct, a large-scale 3D HOI benchmark featuring dataset and methodological advancements. First, we consolidate and standardize 21.81 hours of HOI data from diverse sources, enriching it with detailed textual annotations. Second, we propose a unified optimization framework to enhance data quality by reducing artifacts and correcting hand motions. Leveraging the principle of contact invariance, we maintain human-object relationships while introducing motion variations, expanding the dataset to 30.70 hours. Third, we define six benchmarking tasks and develop a unified HOI generative modeling perspective, achieving state-of-the-art performance. Extensive experiments validate the utility of our dataset as a foundational resource for advancing 3D human-object interaction generation. The dataset will be publicly accessible to support further research in the field.

****

TopoCellGen: Generating Histopathology Cell Topology with a Diffusion Model
Meilong Xu, Saumya Gupta, Xiaoling Hu, Chen Li, Shahira Abousamra, Dimitris Samaras, Prateek Prasanna, Chao Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20979-20989

Accurately modeling multi-class cell topology is crucial in digital pathology, as it provides critical insights into tissue structure and pathology. The synthetic generation of cell topology enables realistic simulations of complex tissue environments, enhances downstream tasks by augmenting training data, aligns more closely with pathologists' domain knowledge, and offers new opportunities for controlling and generalizing the tumor microenvironment. In this paper, we propose a novel approach that integrates topological constraints into a diffusion model to improve the generation of realistic, contextually accurate cell topologies. Our method refines the simulation of cell distributions and interactions, increasing the precision and interpretability of results in downstream tasks such as cell detection and classification. To assess the topological fidelity of generated layouts, we introduce a new metric, Topological Frechet Distance (TopoFD), which overcomes the limitations of traditional metrics like FID in evaluating topological structure. Experimental results demonstrate the effectiveness of our approach in generating multi-class cell layouts that capture intricate topological relationships. Code is available at https://github.com/Melon-Xu/TopoCellGen.

****

Anyattack: Towards Large-scale Self-supervised Adversarial Attacks on Vision-lan

guage Models
Jiaming Zhang, Junhong Ye, Xingjun Ma, Yige Li, Yunfan Yang, Yunhao Chen, Jitao Sang, Dit-Yan Yeung; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19900-19909
Due to their multimodal capabilities, Vision-Language Models (VLMs) have found numerous impactful applications in real-world scenarios. However, recent studies have revealed that VLMs are vulnerable to image-based adversarial attacks. Traditional targeted adversarial attacks require specific targets and labels, limiting their real-world impact. We present AnyAttack, a self-supervised framework that transcends the limitations of conventional attacks through a novel foundation model approach. By pre-training on the massive LAION-400M dataset without label supervision, AnyAttack achieves unprecedented flexibility - enabling any image to be transformed into an attack vector targeting any desired output across different VLMs. This approach fundamentally changes the threat landscape, making adversarial capabilities accessible at an unprecedented scale. Our extensive validation across five open-source VLMs (CLIP, BLIP, BLIP2, InstructBLIP, and MiniGPT-4) demonstrates AnyAttack's effectiveness across diverse multimodal tasks. Most concerning, AnyAttack seamlessly transfers to commercial systems including Google Gemini, Claude Sonnet, Microsoft Copilot and OpenAI GPT, revealing a systemic vulnerability requiring immediate attention.
****

Joint Optimization of Neural Radiance Fields and Continuous Camera Motion from a Monocular Video
Hoang Chuong Nguyen, Wei Mao, Jose M. Alvarez, Miaomiao Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11472-11481
Neural Radiance Fields (NeRF) has demonstrated its superior capability to represent 3D geometry but require accurately precomputed camera poses during training. To mitigate this requirement, existing methods jointly optimize camera poses and NeRF often relying on good pose initialisation or depth priors. However, these approaches struggle in challenging scenarios, such as large rotations, as they map each camera to a world coordinate system. We propose a novel method that eliminates prior dependencies by modeling continuous camera motions as time-dependent angular velocity and velocity. Relative motions between cameras are learned first via velocity integration, while camera poses can be obtained by aggregating such relative motions up to a world coordinate system defined at a single time step within the video. Specifically, accurate continuous camera movements are learned through a time-dependent NeRF, which captures local scene geometry and motion by training from neighboring frames for each time step. The learned motions enable fine-tuning the NeRF to represent the full scene geometry. Experiments on Co3D and Scannet show our approach achieves superior camera pose and depth estimation and comparable novel-view synthesis performance compared to state-of-the-art methods.
****

IRGS: Inter-Reflective Gaussian Splatting with 2D Gaussian Ray Tracing
Chun Gu, Xiaofei Wei, Zixuan Zeng, Yuxuan Yao, Li Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10943-10952
In inverse rendering, accurately modeling visibility and indirect radiance for incident light is essential for capturing secondary effects. Due to the absence of a powerful Gaussian ray tracer, previous 3DGS-based methods have either adopted a simplified rendering equation or used learnable parameters to approximate incident light, resulting in inaccurate material and lighting estimations. To this end, we introduce the inter-reflective Gaussian splatting (IRGS) framework for inverse rendering. To capture inter-reflection, we apply the full rendering equation without simplification and compute incident radiance on the fly using the proposed differentiable 2D Gaussian ray tracing. Additionally, we present an efficient optimization scheme to handle the computational demands of Monte Carlo sampling for rendering equation evaluation. Furthermore, we introduce a novel strategy for querying the indirect radiance of incident light when relighting the optimized scenes. Extensive experiments on multiple standard benchmarks validate the effectiveness of IRGS, demonstrating its capability to accurately model comple

x inter-reflection effects.
****

InterMimic: Towards Universal Whole-Body Control for Physics-Based Human-Object Interactions
Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, Liang-Yan Gui; Proceedings of the Compute r Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12266-12277
Achieving realistic simulations of humans interacting with a wide range of objec ts has long been a fundamental goal. Extending physics-based motion imitation to complex human-object interactions (HOIs) is challenging due to intricate human- object coupling, variability in object geometries, and artifacts in motion captu re data, such as inaccurate contacts and limited hand detail. We introduce Inter Mimic, a framework that enables a single policy to robustly learn from hours of imperfect MoCap data covering diverse full-body interactions with dynamic and va ried objects. Our key insight is to employ a curriculum strategy -- perfect firs t, then scale up. We first train subject-specific teacher policies to mimic, ret arget, and refine motion capture data. Next, we distill these teachers into a st udent policy, with the teachers acting as online experts providing direct superv ision, as well as high-quality references. Notably, we incorporate RL fine-tunin g on the student policy to surpass mere demonstration replication and achieve hi gher-quality solutions. Our experiments demonstrate that InterMimic produces rea listic and diverse interactions across multiple HOI datasets. The learned policy generalizes in a zero-shot manner and seamlessly integrates with kinematic gene rators, elevating the framework from mere imitation to generative modeling of co mplex human-object interactions.
****

Meta-Learning Hyperparameters for Parameter Efficient Fine-Tuning
Zichen Tian, Yaoyao Liu, Qianru Sun; Proceedings of the Computer Vision and Patt ern Recognition Conference (CVPR), 2025, pp. 23037-23047
Training large foundation models from scratch for domain-specific applications i s almost impossible due to data limits and long-tailed distributions -- taking r emote sensing (RS) as an example. Fine-tuning natural image pre-trained models o n RS images is a straightforward solution. To reduce computational costs and imp rove performance on tail classes, existing methods apply parameter-efficient fin e-tuning (PEFT) techniques, such as LoRA and AdaptFormer. However, we observe th at fixed hyperparameters -- such as intra-layer positions, layer depth, and scal ing factors, can considerably hinder PEFT performance, as fine-tuning on RS imag es proves highly sensitive to these settings. To address this, we propose MetaPE FT, a method incorporating adaptive scalers that dynamically adjust module influ ence during fine-tuning. MetaPEFT dynamically adjusts three key factors of PEFT on RS images: module insertion, layer selection, and module-wise learning rates, which collectively control the influence of PEFT modules across the network. We conduct extensive experiments on three transfer-learning scenarios and five dat asets in both RS and natural image domains. The results show that MetaPEFT achie ves state-of-the-art performance in cross-spectral adaptation, requiring only a small amount of trainable parameters and improving tail-class accuracy significa ntly.
****

TriTex: Learning Texture from a Single Mesh via Triplane Semantic Features
Dana Cohen-Bar, Daniel Cohen-Or, Gal Chechik, Yoni Kasten; Proceedings of the Co mputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21403-21413
As 3D content creation continues to grow, transferring semantic textures between 3D meshes remains a significant challenge in computer graphics. While recent me thods leverage text-to-image diffusion models for texturing, they often struggle to preserve the appearance of the source texture during texture transfer. We pr esent TriTex, a novel approach that learns a volumetric texture field from a sin gle textured mesh by mapping semantic features to surface colors. Using an effic ient triplane-based architecture, our method enables semantic-aware texture tran sfer to a novel target mesh. Despite training on just one example, it generalize s effectively to diverse shapes within the same category. Extensive evaluation o n our newly created benchmark dataset shows that TriTex achieves superior textur

e transfer quality and fast inference times compared to existing methods. Our ap
proach advances single-example texture transfer, providing a practical solution
for maintaining visual coherence across related 3D models in applications like g
ame development and simulation.
****

Efficient Test-time Adaptive Object Detection via Sensitivity-Guided Pruning
Kunyu Wang, Xueyang Fu, Xin Lu, Chengjie Ge, Chengzhi Cao, Wei Zhai, Zheng-Jun Z
ha; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)
, 2025, pp. 10577-10586
Continual test-time adaptive object detection (CTTA-OD) aims to online adapt a s
ource pre-trained detector to ever-changing environments during inference under
continuous domain shifts. Most existing CTTA-OD methods prioritize effectiveness
 while overlooking computational efficiency, which is crucial for resource-const
rained scenarios. In this paper, we propose an efficient CTTA-OD method via prun
ing. Our motivation stems from the observation that not all learned source featu
res are beneficial; certain domain-sensitive feature channels can adversely affe
ct target domain performance. Inspired by this, we introduce a sensitivity-guide
d channel pruning strategy that quantifies each channel based on its sensitivity
 to domain discrepancies at both image and instance levels. We apply weighted sp
arsity regularization to selectively suppress and prune these sensitive channels
, focusing adaptation efforts on invariant ones. Additionally, we introduce a st
ochastic channel reactivation mechanism to restore pruned channels, enabling rec
overy of potentially useful features and mitigating the risks of early pruning.
Extensive experiments on three benchmarks show that our method achieves superior
 adaptation performance while reducing computational overhead by 12% in FLOPs co
mpared to the recent SOTA method.
****

A Data-Centric Revisit of Pre-Trained Vision Models for Robot Learning
Xin Wen, Bingchen Zhao, Yilun Chen, Jiangmiao Pang, Xiaojuan Qi; Proceedings of
the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12143-1
2154
Pre-trained vision models (PVMs) are fundamental to modern robotics, yet their o
ptimal configuration remains unclear. Through systematic evaluation, we find tha
t while DINO and iBOT outperform MAE across visuomotor control and perception ta
sks, they struggle when trained on non-(single-)object-centric (NOC) data--a lim
itation strongly correlated with their diminished ability to learn object-centri
c representations. This investigation indicates that the ability to form object-
centric representations from the non-object-centric robotics dataset is the key
to success for PVMs. Motivated by this discovery, we designed SlotMIM, a method
that induces object-centric representations by introducing a semantic bottleneck
 to reduce the number of prototypes to encourage the emergence of objectness as
well as cross-view consistency regularization for encouraging multiview invarian
ce. Our experiments encompass pre-training on object-centric, scene-centric, web
-crawled, and ego-centric data. Across all settings, our approach learns transfe
rrable representations and achieves significant improvements over prior work in
image recognition, scene understanding, and robot learning evaluations. When sca
led up with million-scale datasets, our method also demonstrates superior data e
fficiency and scalability. Our code and models are publicly available at https:/
/github.com/CVMI-Lab/SlotMIM.
****

Visual Agentic AI for Spatial Reasoning with a Dynamic API
Damiano Marsili, Rohun Agrawal, Yisong Yue, Georgia Gkioxari; Proceedings of the
 Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19446-1945
5
Visual reasoning -- the ability to interpret the visual world -- is crucial for
embodied agents that operate within three-dimensional scenes. Progress in AI has
 led to vision and language models capable of answering questions from images. H
owever, their performance declines when tasked with 3D spatial reasoning. To tac
kle the complexity of such reasoning problems, we introduce an agentic program s
ynthesis approach where LLM agents collaboratively generate a Pythonic API with

new functions to solve common subproblems. Our method overcomes limitations of p
rior approaches that rely on a static, human-defined API, allowing it to handle
a wider range of queries. To assess AI capabilities for 3D understanding, we int
roduce a new benchmark of queries involving multiple steps of grounding and infe
rence. We show that our method outperforms prior zero-shot models for visual rea
soning in 3D and empirically validate the effectiveness of our agentic framework
for 3D spatial reasoning tasks. Project website: https://glab-caltech.github.io
/vadar/
****

TAMT: Temporal-Aware Model Tuning for Cross-Domain Few-Shot Action Recognition
Yilong Wang, Zilin Gao, Qilong Wang, Zhaofeng Chen, Peihua Li, Qinghua Hu; Proce
edings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, p
p. 3449-3459
Going beyond few-shot action recognition (FSAR), cross-domain FSAR (CDFSAR) has
attracted recent research interests by solving the domain gap lying in source-to
-target transfer learning. Existing CDFSAR methods mainly focus on joint trainin
g of source and target data to mitigate the side effect of domain gap. However,
such kind of methods suffer from two limitations: First, pair-wise joint trainin
g requires retraining deep models in case of one source data and multiple target
ones, which incurs heavy computation cost, especially for large source and smal
l target data. Second, pre-trained models after joint training are adopted to ta
rget domain in a straightforward manner, hardly taking full potential of pre-tra
ined models and then limiting recognition performance. To overcome above limitat
ions, this paper proposes a simple yet effective baseline, namely Temporal-Aware
Model Tuning (TAMT) for CDFSAR. Specifically, our TAMT involves a decoupled par
adigm by performing pre-training on source data and fine-tuning target data, whi
ch avoids retraining for multiple target data with single source. To effectively
and efficiently explore the potential of pre-trained models in transferring to
target domain, our TAMT proposes a Hierarchical Temporal Tuning Network (HTTN),
whose core involves local temporal-aware adapters (TAA) and a global temporal-aw
are moment tuning (GTMT). Particularly, TAA learns few parameters to recalibrate
the intermediate features of frozen pre-trained models, enabling efficient adap
tation to target domains. Furthermore, GTMT helps to generate powerful video rep
resentations, improving match performance on the target domain. Experiments on s
everal widely used video benchmarks show our TAMT outperforms the recently propo
sed counterparts by 13% 31%, achieving new state-of-the-art CDFSAR results.
****

Feature Spectrum Learning for Remote Sensing Change Detection
Qi Zang, Dong Zhao, Shuang Wang, Dou Quan, Zhun Zhong; Proceedings of the Comput
er Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12647-12657
Change detection (CD) holds significant implications for Earth observation, in w
hich pseudo-changes between bitemporal images induced by imaging environmental f
actors are key challenges. Existing methods mainly regard pseudo-changes as a ki
nd of style shift and alleviate it by transforming bitemporal images into the sa
me style using generative adversarial networks (GANs). Nevertheless, their effor
ts are limited by the complexity of optimizing GANs and the absence of guidance
from physical properties. This paper finds that the spectrum transformation (ST)
has the potential to mitigate pseudo-changes by aligning in the frequency domai
n carrying the style. However, the benefit of ST is largely constrained by two d
rawbacks: 1) limited transformation space and 2) inefficient parameter search. T
o address these limitations, we propose a Feature Spectrum learning (FeaSpect) t
hat adaptively eliminate pseudo-changes in the latent space. For the drawback 1)
, FeaSpect directs the transformation towards style-aligned discriminative featu
res via feature spectrum transformation (FST). For the drawback 2), FeaSpect all
ows FST to be trainable, efficiently discovering optimal parameters via extracti
on box with adaptive attention and extraction box with learnable strides. Extens
ive experiments on challenging datasets demonstrate that our method remarkably o
utperforms existing methods and achieves a commendable trade-off between accurac
y and efficiency. Importantly, our method can be easily injected into other fram
eworks, achieving consistent improvements.

****
BioX-CPath: Biologically-driven Explainable Diagnostics for Multistain IHC Computational Pathology
Amaya Gallagher-Syed, Henry Senior, Omnia Alwazzan, Elena Pontarini, Michele Bombardieri, Costantino Pitzalis, Myles J. Lewis, Michael R. Barnes, Luca Rossi, Gregory Slabaugh; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10372-10383
The development of biologically interpretable and explainable models remains a key challenge in computational pathology, particularly for multistain immunohistochemistry (IHC) analysis. We present BioX-CPath, an explainable graph neural network architecture for whole slide image (WSI) classification that leverages both spatial and semantic features across multiple stains. At its core, BioXCPath introduces a novel Stain-Aware Attention Pooling (SAAP) module that generates biologically meaningful, stain-aware patient embeddings. Our approach achieves state-of-the-art performance on both Rheumatoid Arthritis and Sjogren's Disease multistain datasets. Beyond performance metrics, BioX-CPath provides interpretable insights through stain attention scores, entropy measures, and stain interaction scores, that permit measuring model alignment with known pathological mechanisms. This biological grounding, combined with strong classification performance, makes BioX-CPath particularly suitable for clinical applications where interpretability is key. Source code and documentation can be found at: https://github.com/AmayaGS/BioX-CPath.
****
DriveDreamer4D: World Models Are Effective Data Machines for 4D Driving Scene Representation
Guosheng Zhao, Chaojun Ni, Xiaofeng Wang, Zheng Zhu, Xueyang Zhang, Yida Wang, Guan Huang, Xinze Chen, Boyuan Wang, Youyi Zhang, Wenjun Mei, Xingang Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12015-12026
Closed-loop simulation is essential for advancing end-to-end autonomous driving systems. Contemporary sensor simulation methods, such as NeRF and 3DGS, rely predominantly on conditions closely aligned with training data distributions, which are largely confined to forward-driving scenarios. Consequently, these methods face limitations when rendering complex maneuvers (e.g., lane change, acceleration, deceleration). Recent advancements in autonomous-driving world models have demonstrated the potential to generate diverse driving videos. However, these approaches remain constrained to 2D video generation, inherently lacking the spatiotemporal coherence required to capture intricacies of dynamic driving environments.In this paper, we introduce DriveDreamer4D, which enhances 4D driving scene representation leveraging world model priors. Specifically, we utilize the world model as a data machine to synthesize novel trajectory videos, where structured conditions are explicitly leveraged to control the spatial-temporal consistency of traffic elements. Besides, the cousin data training strategy is proposed to facilitate merging real and synthetic data for optimizing 4DGS. To our knowledge, DriveDreamer4D is the first to utilize video generation models for improving 4D reconstruction in driving scenarios.Experimental results reveal that DriveDreamer4D significantly enhances generation quality under novel trajectory views, achieving a relative improvement in FID by 32.1%, 46.4%, and 16.3% compared to PVG, S3Gaussian, and Deformable-GS. Moreover, DriveDreamer4D markedly enhances the spatiotemporal coherence of driving agents, which is verified by a comprehensive user study and the relative increases of 22.6%, 43.5%, and 15.6% in the NTA-IoU metric.
****
LoKi: Low-dimensional KAN for Efficient Fine-tuning Image Models
Xuan Cai, Renjie Pan, Hua Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14869-14880
'Pre-training + fine-tuning' has been widely used in various downstream tasks. Parameter-efficient fine-tuning (PEFT) has demonstrated higher efficiency and promising performance comapred to traditional full-tuning. The widely used adapter-based and prompt-based methods in PEFT can be uniformly represented as adding an

MLP structure to the pre-trained model. These methods are prone to over-fitting in downstream tasks, due to the difference in data scale and distribution. To address this issue, we propose a new adapter-based PEFT module, i.e., LoKi, which consists of an encoder, a learnable activation layer, and a decoder. To maintain the simplicity of LoKi, we use single-layer linear networks for the encoder and decoder, and for the learnable activation layer, we use a Kolmogorov-Arnold Network (KAN) with the minimal number of layers (only 2 KAN linear layers). With a bottleneck rate much lower than that of Adapter, LoKi is equipped with fewer parameters (only half of Adapter) and eliminates slow training speed and high memory usage of KAN. We conduct extensive experiments on LoKi under image classification and video action recognition across 9 datasets. LoKi demonstrates highly competitive generalization performance compared to other PEFT methods with fewer tunable parameters, ensuring both effectiveness and efficiency.
****

Dr. Splat: Directly Referring 3D Gaussian Splatting via Direct Language Embedding Registration

Kim Jun-Seong, GeonU Kim, Kim Yu-Ji, Yu-Chiang Frank Wang, Jaesung Choe, Tae-Hyun Oh; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14137-14146

We introduce Dr. Splat, a novel approach for open-vocabulary 3D scene understanding leveraging 3D Gaussian Splatting. Unlike existing language-embedded 3DGS methods, which rely on a rendering process, our method directly associates language-aligned CLIP embeddings with 3D Gaussians for holistic 3D scene understanding. The key of our method is a language feature registration technique where CLIP embeddings are assigned to the dominant Gaussians intersected by each pixel-ray. Moreover, we integrate Product Quantization (PQ) trained on general large scale image data to compactly represent embeddings without per-scene optimization. Experiments demonstrate that our approach significantly outperforms existing approaches in 3D perception benchmarks, such as open-vocabulary 3D semantic segmentation, 3D object localization, and 3D object selection tasks.
****

Wavelet and Prototype Augmented Query-based Transformer for Pixel-level Surface Defect Detection

Feng Yan, Xiaoheng Jiang, Yang Lu, Jiale Cao, Dong Chen, Mingliang Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23860-23869

As an important part of intelligent manufacturing, pixel-level surface defect detection (SDD) aims to locate defect areas through mask prediction. Previous methods adopt the image-independent static convolution to indiscriminately classify per-pixel features for mask prediction, which leads to suboptimal results for some challenging scenes such as weak defects and cluttered backgrounds. In this paper, inspired by query-based methods, we propose a Wavelet and Prototype Augmented Query-based Transformer (WPFormer) for surface defect detection. Specifically, a set of dynamic queries for mask prediction is updated through the dual-domain transformer decoder. Firstly, a Wavelet-enhanced Cross-Attention (WCA) is proposed, which aggregates meaningful high- and low-frequency information of image features in the wavelet domain to refine queries. WCA enhances the representation of high-frequency components by capturing multi-scale relationships between different frequency components, enabling queries to focus more on defect details. Secondly, a Prototype-guided Cross-Attention (PCA) is proposed to refine queries through meta-prototypes in the spatial domain. The prototypes aggregate semantically meaningful tokens from image features, facilitating queries to aggregate crucial defect information under the cluttered backgrounds. Extensive experiments on three defect detection datasets (i.e., ESDIs-SOD, CrackSeg9k, and ZJU-Leaper) demonstrate that the proposed method achieves state-of-the-art performance in defect detection. The code will be available at https://github.com/yfhdm/WPFormer.
****

GauCho: Gaussian Distributions with Cholesky Decomposition for Oriented Object Detection

José Henrique Lima Marques, Jeffri Murrugarra-Llerena, Claudio R. Jung; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3593-3602

Oriented Object Detection (OOD) has received increased attention in the past years, being a suitable solution for detecting elongated objects in remote sensing analysis. In particular, using regression loss functions based on Gaussian distributions has become attractive since they yield simple and differentiable terms. However, existing solutions are still based on regression heads that produce Oriented Bounding Boxes (OBBs), and the known problem of angular boundary discontinuity persists. In this work, we propose a regression head for OOD that directly produces Gaussian distributions based on the Cholesky matrix decomposition. The proposed head, named Gaucho, theoretically mitigates the boundary discontinuity problem and is fully compatible with recent Gaussian-based regression loss functions. Furthermore, we advocate using Oriented Ellipses (OEs) to represent oriented objects, which relates to GauCho through a bijective function and alleviates the encoding ambiguity problem for circular objects. Our experimental results show that GauCho can be a viable alternative to the traditional OBB head, achieving results comparable to or better than state-of-the-art detectors for the challenging dataset DOTA.

****

Alignment, Mining and Fusion: Representation Alignment with Hard Negative Mining and Selective Knowledge Fusion for Medical Visual Question Answering

Yuanhao Zou, Zhaozheng Yin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29623-29633

Medical Visual Question Answering (Med-VQA) is a challenging task that requires a deep understanding of both medical images and textual questions. Although recent works leveraging Medical Vision-Language Pre-training (Med-VLP) have shown strong performance on the Med-VQA task, there is still no unified solution for modality alignment, and the issue of hard negatives remains under-explored. Additionally, commonly used knowledge fusion techniques for Med-VQA may introduce irrelevant information. In this work, we propose a framework to address these challenges through three key contributions: (1) a unified solution for heterogeneous modality alignments across multiple levels, modalities, views, and stages, leveraging methods such as contrastive learning and optimal transport theory; (2) a hard negative mining method that employs soft labels for multi-modality alignments and enforces the hard negative pair discrimination; and (3) a Gated Cross-Attention Module for Med-VQA that integrates the answer vocabulary as prior knowledge and select relevant information from it. Our framework outperforms the previous state-of-the-art on widely used Med-VQA datasets like RAD-VQA, SLAKE, PathVQA and VQA-2019. The code will be publicly available.

****

No Thing, Nothing: Highlighting Safety-Critical Classes for Robust LiDAR Semantic Segmentation in Adverse Weather

Junsung Park, Hwijeong Lee, Inha Kang, Hyunjung Shim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6690-6699

Existing domain generalization methods for LiDAR semantic segmentation under adverse weather struggle to accurately predict "things" categories compared to "stuff" categories. In typical driving scenes, "things" categories can be dynamic and associated with higher collision risks, making them crucial for safe navigation and planning. Recognizing the importance of "things" categories, we identify their performance drop as a serious bottleneck in existing approaches. We observed that adverse weather induces degradation of semantic-level features and both corruption of local features, leading to a misprediction of "things" as "stuff". To mitigate these corruptions, we suggests our method, NTN - segmeNt Things for No-accident. To address semantic-level feature corruption, we bind each point feature to its superclass, preventing the misprediction of things classes into visually dissimilar categories. Additionally, to enhance robustness against local corruption caused by adverse weather, we define each LiDAR beam as a local region and propose a regularization term that aligns the clean data with its corrupted counterpart in feature space. NTN achieves state-of-the-art performance with a

+2.6 mIoU gain on the SemanticKITTI-to-SemanticSTF benchmark and +7.9 mIoU on th
e SemanticPOSS-to-SemanticSTF benchmark. Notably, NTN achieves a +4.8 and +7.9 m
IoU improvement on "things" classes, respectively, highlighting its effectivenes
s.
****

Mind the Gap: Detecting Black-box Adversarial Attacks in the Making through Quer
y Update Analysis
Jeonghwan Park, Niall McLaughlin, Ihsen Alouani; Proceedings of the Computer Vis
ion and Pattern Recognition Conference (CVPR), 2025, pp. 10235-10243
Adversarial attacks remain a significant threat that can jeopardize the integrit
y of Machine Learning (ML) models. In particular, query-based black-box attacks
can generate malicious noise without having access to the victim model's archite
cture, making them practical in real-world contexts. The community has proposed
several defenses against adversarial attacks, only to be broken by more advanced
 and adaptive attack strategies. In this paper, we propose a framework that det
ects if an adversarial noise instance is being generated. Unlike existing statef
ul defenses that detect adversarial noise generation by monitoring the input spa
ce, our approach learns adversarial patterns in the input update similarity spac
e. In fact, we propose to observe a new metric called Delta Similarity (DS), whi
ch we show it captures more efficiently the adversarial behavior.We evaluate our
 approach against 8 state-of-the-art attacks, including adaptive attacks, where
the adversary is aware of the defense and tries to evade detection. We find that
 our approach is significantly more robust than existing defenses both in terms
of specificity and sensitivity.
****

Consistent Normal Orientation for 3D Point Clouds via Least Squares on Delaunay
Graph
Rao Fu, Jianmin Zheng, Liang Yu; Proceedings of the Computer Vision and Pattern
Recognition Conference (CVPR), 2025, pp. 16932-16942
The orientation of surface normals in 3D point cloud is a fundamental problem in
 computer vision and graphics. Determining a globally consistent orientation sol
ely from the point cloud is however challenging due to the global scope of the p
roblem and the discrete nature of point cloud, particularly in the presence of n
oise, outliers, holes, thin structures, and complex topologies. This paper prese
nts an efficient, robust, and global algorithm for generating consistent normal
orientation of a dense 3D point cloud. The basic idea is to transform the origin
al binary normal orientation problem to finding a relaxed sign field on a Delaun
ay graph, which can be achieved by solving a sparse linear system. The Delaunay
graph is constructed by triangulating a level set of an implicit function define
d from the input point cloud. The shape diameter function is estimated to serve
as a prior for determining an appropriate level value such that the level set im
plicitly defines the inner and outer shells enclosing the input point clouds. As
 such, our algorithm leverages the strengths of the shape diameter function, Del
aunay triangulation, and the least-square techniques, making the underlying proc
esses take both geometry and topology into consideration, and thus provides an e
fficient and robust solution for handling point clouds with complicated geometry
 and topology. Extensive experiments on various shapes with noise and outliers c
onfirm the effectiveness and robustness of our algorithm.
****

GaussianWorld: Gaussian World Model for Streaming 3D Occupancy Prediction
Sicheng Zuo, Wenzhao Zheng, Yuanhui Huang, Jie Zhou, Jiwen Lu; Proceedings of th
e Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6772-6781
3D occupancy prediction is important for autonomous driving due to its comprehen
sive perception of the surroundings. To incorporate sequential inputs, most exis
ting methods fuse representations from previous frames to infer the current 3D o
ccupancy. However, they fail to consider the continuity of driving scenarios and
 ignore the strong prior provided by the evolution of 3D scenes (e.g., only dyna
mic objects move). In this paper, we propose a world-modelbased framework to exp
loit the scene evolution for perception. We reformulate 3D occupancy prediction
as a 4D occupancy forecasting problem conditioned on the current sensor input. W

e decompose the scene evolution into three factors: 1) ego motion alignment of s
tatic scenes; 2) local movements of dynamic objects; and 3) completion of newly-
observed scenes. We then employ a Gaussian world model (GaussianWorld) to explic
itly exploit these priors and infer the scene evolution in the 3D Gaussian space
 considering the current RGB observation. We evaluate the effectiveness of our f
ramework on the widely used nuScenes dataset. Our GaussianWorld improves the per
formance of the single-frame counterpart by over 2% in mIoU without introducing
additional computations.
****

ICP: Immediate Compensation Pruning for Mid-to-high Sparsity
Xin Luo, Xueming Fu, Zihang Jiang, S. Kevin Zhou; Proceedings of the Computer Vi
sion and Pattern Recognition Conference (CVPR), 2025, pp. 9487-9496
The increasing adoption of large-scale models under 7 billion parameters in both
 language and vision domains enables inference tasks on a single consumer-grade
GPU but makes fine-tuning models of this scale, especially 7B models, challengin
g. This limits the applicability of pruning methods that require full fine-tunin
g. Meanwhile, pruning methods that do not require fine-tuning perform well at lo
w sparsity levels (10%-50%) but struggle at mid-to-high sparsity levels (50%-70%
), where the error behaves equivalently to that of semi-structured pruning. To a
ddress these issues, this paper introduces ICP, which finds a balance between fu
ll fine-tuning and zero fine-tuning. First, Sparsity Rearrange is used to reorga
nize the predefined sparsity levels, followed by Block-wise Compensate Pruning,
which alternates pruning and compensation on the model's backbone, fully utilizi
ng inference results while avoiding full model fine-tuning. Experiments show tha
t ICP improves performance at mid-to-high sparsity levels compared to baselines,
 with only a slight increase in pruning time and no additional peak memory overh
ead.
****

VinaBench: Benchmark for Faithful and Consistent Visual Narratives
Silin Gao, Sheryl Mathew, Li Mi, Sepideh Mamooler, Mengjie Zhao, Hiromi Wakaki,
Yuki Mitsufuji, Syrielle Montariol, Antoine Bosselut; Proceedings of the Compute
r Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2870-2879
Visual narrative generation transforms textual narratives into sequences of imag
es illustrating the content of the text. However, generating visual narratives t
hat are faithful to the input text and self-consistent across generated images r
emains an open challenge, due to the lack of knowledge constraints used for plan
ning the stories. In this work, we propose a new benchmark, VinaBench, to addres
s this challenge. Our benchmark annotates the underlying commonsense and discour
se constraints in visual narrative samples, offering systematic scaffolds for le
arning the implicit strategies of visual storytelling. Based on the incorporated
 narrative constraints, we further propose novel metrics to closely evaluate the
 consistency of generated narrative images and the alignment of generations with
 the input textual narrative. Our results across three generative vision models
demonstrate that learning with VinaBench's knowledge constraints effectively imp
roves the faithfulness and cohesion of generated visual narratives.
****

ATA: Adaptive Transformation Agent for Text-Guided Subject-Position Variable Bac
kground Inpainting
Yizhe Tang, Zhimin Sun, Yuzhen Du, Ran Yi, Guangben Lu, Teng Hu, Luying Li, Lizh
uang Ma, Fangyuan Zou; Proceedings of the Computer Vision and Pattern Recognitio
n Conference (CVPR), 2025, pp. 18335-18345
Image inpainting aims to fill the missing region of an image.Recently, there has
 been a surge of interest in foreground-conditioned background inpainting, a sub
-task that fills the background of an image while the foreground subject and ass
ociated text prompt are provided.Existing background inpainting methods typicall
y strictly preserve the subject's original position from the source image,result
ing in inconsistencies between the subject and the generated background.To addre
ss this challenge, we propose a new task, the "Text-Guided Subject-Position Vari
able Background Inpainting", which aims to dynamically adjust the subject positi
on to achieve a harmonious relationship between the subject andthe inpainted bac

kground, and propose the Adaptive Transformation Agent (A^\text T A) for this ta sk.Firstly, we design a PosAgent Block that adaptively predicts an appropriate d isplacement based on given features to achieve variable subject-position.Secondl y, we design the Reverse Displacement Transform (RDT) module, which arranges mul tiple PosAgent blocks in a reverse structure, to transform hierarchical feature maps from deep to shallow based on semantic information.Thirdly, we equip A^\tex t T A with a Position Switch Embedding to control whether the subject's position in the generated image is adaptively predicted or fixed.Extensive comparative e xperiments validate the effectiveness of our A^\text T A approach, which not onl y demonstrates superior inpainting capabilities in subject-position variable inp ainting, but also ensures good performance on subject-position fixed inpainting.

****

Optimizing for the Shortest Path in Denoising Diffusion Model

Ping Chen, Xingpeng Zhang, Zhaoxiang Liu, Huan Hu, Xiang Liu, Kai Wang, Min Wang , Yanlin Qian, Shiguo Lian; Proceedings of the Computer Vision and Pattern Recog nition Conference (CVPR), 2025, pp. 18021-18030

In this research, we propose a novel denoising diffusion model based on shortest -path modeling that optimizes residual propagation to enhance both denoising eff iciency and quality. Drawing on Denoising Diffusion Implicit Models (DDIM) and i nsights from graph theory, our model, termed the Shortest Path Diffusion Model (ShortDF), treats the denoising process as a shortest-path problem aimed at minim izing reconstruction error. By optimizing the initial residuals, we improve the efficiency of the reverse diffusion process and the quality of the generated sam ples. Extensive experiments on multiple standard benchmarks demonstrate that Sho rtDF significantly reduces diffusion time (or steps) while enhancing the visual fidelity of generated samples compared to prior methods. This work, we suppose, paves the way for interactive diffusion-based applications and establishes a fou ndation for rapid data generation. Code is available at https://github.com/Unico mAI/ShortDF.

****

Antidote: A Unified Framework for Mitigating LVLM Hallucinations in Counterfactu al Presupposition and Object Perception

Yuanchen Wu, Lu Zhang, Hang Yao, Junlong Du, Ke Yan, Shouhong Ding, Yunsheng Wu, Xiaoqiang Li; Proceedings of the Computer Vision and Pattern Recognition Confer ence (CVPR), 2025, pp. 14646-14656

Large Vision-Language Models (LVLMs) have achieved impressive results across var ious cross-modal tasks. However, hallucinations, i.e., the models generating cou nterfactual responses, remain a challenge. Though recent studies have attempted to alleviate object perception hallucinations, they focus on the models' respons e generation, and overlook the task question itself. This paper discusses the vu lnerability of LVLMs in solving counterfactual presupposition questions (CPQs), where the models are prone to accept the presuppositions of counterfactual objec ts and produce severe hallucinatory responses. To this end, we introduce "Antido te", a unified, synthetic data-driven post-training framework for mitigating bot h types of hallucination above. It leverages synthetic data to incorporate factu al priors into questions to achieve self-correction, and decouples the mitigatio n process into a preference optimization problem. Furthermore, we construct "CP-Bench", a novel benchmark to evaluate LVLMs' ability to correctly handle CPQs an d produce factual responses. Applied to the LLaVA series, Antidote can simultane ously enhance performance on CP-Bench by over 50%, POPE by 1.8-3.3%, and CHAIR & SHR by 30-50%, all without relying on external supervision from stronger LVLMs or human feedback and without introducing noticeable catastrophic forgetting iss ues.

****

Language-Guided Audio-Visual Learning for Long-Term Sports Assessment

Huangbiao Xu, Xiao Ke, Huanqi Wu, Rui Xu, Yuezhou Li, Wenzhong Guo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2396 7-23977

Long-term sports assessment is a challenging task in video understanding since i t requires judging complex movement variations and action-music coordination. Ho

wever, there is no direct correlation between the diverse background music and m
ovements in sporting events. Previous works require a large number of model para
meters to learn potential associations between actions and music. To address thi
s issue, we propose a language-guided audio-visual learning (MLAVL) framework th
at models "audio-action-visual" correlations guided by low-cost language modalit
y. In our framework, multidimensional domain-based actions form action knowledge
 graphs, motivating audio-visual modalities to focus on task-relevant actions. W
e further design a shared-specific context encoder to integrate deep multimodal
semantics, and an audio-visual cross-modal fusion module to evaluate action-musi
c consistency. To match the sport's rules, we then propose a dual-branch prompt-
guided grading module to weigh both visual and audio-visual performance. Extensi
ve experiments demonstrate that our approach achieves state-of-the-art on four p
ublic long-term sports benchmarks while maintaining low parameters. Our code is
available at https://github.com/XuHuangbiao/MLAVL.
****

Dynamic Pseudo Labeling via Gradient Cutting for High-Low Entropy Exploration
Jae Hyeon Park, Joo Hyeon Jeon, Jae Yun Lee, Sangyeon Ahn, Min Hee Cha, Min Geol
 Kim, Hyeok Nam, Sung In Cho; Proceedings of the Computer Vision and Pattern Rec
ognition Conference (CVPR), 2025, pp. 20602-20611
This study addresses the limitations of existing dynamic pseudo-labeling (DPL) t
echniques, which often utilize static or dynamic thresholds for confident sample
 selection. The existing methods fail to capture the non-linear relationship bet
ween task accuracy and model confidence, particularly in the context of overconf
idence. This can limit the model's learning opportunities for high entropy sampl
es that significantly influence a model's generalization ability. To solve this,
 we propose a novel gradient pass-based DPL technique that incorporates the high
-entropy samples, which are typically overlooked. Our approach introduces two cl
assifiers--low gradient pass (LGP) and high gradient pass (HGP)--to derive over-
 and under-confident dynamic thresholds that indicate the class-wise overconfide
nce acceleration, respectively. By combining the under- and over-confident state
s from the GP classifiers, we create a more adaptive and accurate PL method. Our
 main contributions highlight the importance of considering both low and high-co
nfidence samples in enhancing the model's robustness and generalization for impr
oved PL performance.
****

VODiff: Controlling Object Visibility Order in Text-to-Image Generation
Dong Liang, Jinyuan Jia, Yuhao Liu, Zhanghan Ke, Hongbo Fu, Rynson W. H. Lau; Pr
oceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025
, pp. 18379-18389
Recent advancements in diffusion models have significantly enhanced the performa
nce of text-to-image models in image synthesis. To enable control over the the s
patial locations of the generated objects,diffusion-based methods typically util
izeobject layout as an auxiliary input. However, we observe that this approach t
reats all objects as being on the same layer and neglect their visibility order,
 leading to the synthesis of overlapping objects with incorrect occlusions.To ad
dress this limitation, we introduce in this paper a new training-free framework
that considers object visibility order explicitly and allows users to place over
lapping objects in a stack of layers. Our framework consists of two visibility-b
ased designs. First, we propose a novel Sequential Denoising Process (SDP) to di
vide the whole image generation into multiple stages for different objects, each
 stage primarily focuses on an object. Second, we propose a novel Visibility-Ord
er-Aware (VOA) Loss to transform the layout and occlusion constraints into an at
tention map optimization process to improve the accuracy of synthesizing object
occlusions in complex scenes. By merging these two novel components, our framewo
rk, dubbed VODiff, enables the generation of photorealistic images that satisfy
user-specified spatial constraints and object occlusion relationships. In additi
on, we introduce VOBench, a diverse benchmark dataset containing 200 curated sam
ples, each with a reference image, text prompts, object visibility orders and la
yout maps. We conduct extensive evaluations on this dataset to demonstrate the s
uperiority of our approach.

****
Dual Diffusion for Unified Image Generation and Understanding
Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, Peng Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2779-2790
Diffusion models have gained tremendous success in text-to-image generation, yet still struggle with visual understanding tasks, an area dominated by autoregressive vision-language models. We propose a large-scale and fully end-to-end diffusion model for multi-modal understanding and generation that significantly improves on existing diffusion-based multimodal models, and is the first of its kind to support the full suite of vision-language modeling capabilities. Inspired by the multimodal diffusion transformer (MM-DiT) and recent advances in discrete diffusion language modeling, we leverage a cross-modal maximum likelihood estimation framework that simultaneously trains the conditional likelihoods of both images and text jointly under a single loss function, which is back-propagated through both branches of the diffusion transformer. The resulting model is highly flexible and capable of a wide range of tasks including image generation, captioning, and visual question answering. Our model attained competitive performance compared to recent unified image understanding and generation models, demonstrating the potential of multimodal diffusion modeling as a promising alternative to autoregressive next-token prediction models.
****
WeakMCN: Multi-task Collaborative Network for Weakly Supervised Referring Expression Comprehension and Segmentation
Silin Cheng, Yang Liu, Xinwei He, Sebastien Ourselin, Lei Tan, Gen Luo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9175-9185
Weakly supervised referring expression comprehension (WREC) and segmentation (WRES) aim to learn object grounding based on a given expression using weak supervision signals like image-text pairs. While these tasks have traditionally been modeled separately, we argue that they can benefit from joint learning in a multi-task framework. To this end, we propose WeakMCN, a novel multi-task collaborative network that effectively combines WREC and WRES with a dual-branch architecture. Specifically, the WREC branch is formulated as anchor-based contrastive learning, which also acts as a teacher to supervise the WRES branch. In WeakMCN, we propose two innovative designs to facilitate multi-task collaboration, namely Dynamic Visual Feature Enhancement (DVFE) and Collaborative Consistency Module (CCM). DVFE dynamically combines various pre-trained visual knowledge to meet different task requirements, while CCM promotes cross-task consistency from the perspective of optimization. Extensive experimental results on three popular REC and RES benchmarks, i.e., RefCOCO, RefCOCO+, and RefCOCOg, consistently demonstrate performance gains of WeakMCN over state-of-the-art single-task alternatives, e.g., up to 3.91% and 13.11% on RefCOCO for WREC and WRES tasks, respectively. Furthermore, experiments also validate the strong generalization ability of WeakMCN in both semi-supervised REC and RES settings against existing methods, e.g., +8.94% for semi-REC and +7.71% for semi-RES on 1% RefCOCO.
****
CAD-Llama: Leveraging Large Language Models for Computer-Aided Design Parametric 3D Model Generation
Jiahao Li, Weijian Ma, Xueyang Li, Yunzhong Lou, Guichun Zhou, Xiangdong Zhou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18563-18573
Recently, Large Language Models (LLMs) have achieved significant success, prompting increased interest in expanding their generative capabilities beyond general text into domain-specific areas. This study investigates the generation of parametric sequences for computer-aided design (CAD) models using LLMs. This endeavor represents an initial step towards creating parametric 3D shapes with LLMs, as CAD model parameters directly correlate with shapes in three-dimensional space. Despite the formidable generative capacities of LLMs, this task remains challenging, as these models neither encounter parametric sequences during their pretra

ining phase nor possess direct awareness of 3D structures. To address this, we p
resent CAD-Llama, a framework designed to enhance pretrained LLMs for generating
 parametric 3D CAD models. Specifically, we develop a hierarchical annotation pi
peline and a code-like format to translate parametric 3D CAD command sequences i
nto Structured Parametric CAD Code (SPCC), incorporating hierarchical semantic d
escriptions. Furthermore, we propose an adaptive pretraining approach utilizing
 SPCC, followed by an instruction tuning process aligned with CAD-specific guidel
ines. This methodology aims to equip LLMs with the spatial knowledge inherent in
 parametric sequences. Experimental results demonstrate that our framework signi
ficantly outperforms prior autoregressive methods and existing LLM baselines.
****

Leveraging SD Map to Augment HD Map-based Trajectory Prediction
Zhiwei Dong, Ran Ding, Wei Li, Peng Zhang, Guobin Tang, Jia Guo; Proceedings of
 the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17219-1
7228
Latest trajectory prediction models in real-world autonomous driving systems oft
en rely on online High-Definition (HD) maps to understand the road environment.H
owever, online HD maps suffer from perception errors and feature redundancy, whi
ch hinder the performance of HD map-based trajectory prediction models.To addres
s these issues, we introduce a framework, termed SD map-Augmented Trajectory Pre
diction (SATP), which leverages Standard-Definition (SD) maps to enhance HD map-
based trajectory prediction models.First, we propose an SD-HD fusion approach to
 leverage SD maps across the diverse range of HD map-based trajectory prediction
 models. Second, we design a novel AlignNet to align the SD map with the HD map,
 further improving the effectiveness of SD maps. Experiments on real-world auton
omous driving benchmarks demonstrate that SATP not only improves the performance
 of HD map-based trajectory prediction up to 25% in real-world scenarios using o
nline HD maps but also brings benefits in ideal scenarios with ground-truth HD m
aps.
****

4DGC: Rate-Aware 4D Gaussian Compression for Efficient Streamable Free-Viewpoint
 Video
Qiang Hu, Zihan Zheng, Houqiang Zhong, Sihua Fu, Li Song, Xiaoyun Zhang, Guangta
o Zhai, Yanfeng Wang; Proceedings of the Computer Vision and Pattern Recognition
 Conference (CVPR), 2025, pp. 875-885
3D Gaussian Splatting (3DGS) has substantial potential for enabling photorealist
ic Free-Viewpoint Video (FVV) experiences. However, the vast number of Gaussians
 and their associated attributes poses significant challenges for storage and tr
ansmission. Existing methods typically handle dynamic 3DGS representation and co
mpression separately, neglecting motion information and the rate-distortion (RD)
 trade-off during training, leading to performance degradation and increased mod
el redundancy. To address this gap, we propose 4DGC, a novel rate-aware 4D Gauss
ian compression framework that significantly reduces storage size while maintai
ning superior RD performance for FVV. Specifically, 4DGC introduces a motion-awa
re dynamic Gaussian representation that utilizes a compact motion grid combined
with sparse compensated Gaussians to exploit inter-frame similarities. This repr
esentation effectively handles large motions, preserving quality and reducing te
mporal redundancy. Furthermore, we present an end-to-end compression scheme that
 employs differentiable quantization and a tiny implicit entropy model to compre
ss the motion grid and compensated Gaussians efficiently. The entire framework i
s jointly optimized using a rate-distortion trade-off. Extensive experiments dem
onstrate that 4DGC supports variable bitrates and consistently outperforms exist
ing methods in RD performance across multiple datasets.
****

ONDA-Pose: Occlusion-Aware Neural Domain Adaptation for Self-Supervised 6D Objec
t Pose Estimation
Tao Tan, Qiulei Dong; Proceedings of the Computer Vision and Pattern Recognition
 Conference (CVPR), 2025, pp. 16829-16838
Self-supervised 6D object pose estimation has received increasing attention in c
omputer vision recently. Some typical works in literature attempt to translate t

he synthetic images with object pose labels generated by object CAD models into the real domain, and then use the translated data for training. However, their performance is generally limited, since (i) there still exists a domain gap between the translated images and the real images and (ii) the translated images cannot sufficiently reflect occlusions that exist in many real images. To address these problems, we propose an Occlusion-Aware Neural Domain Adaptation method for self-supervised 6D object Pose estimation, called ONDA-Pose. The proposed method comprises three main steps. Firstly, by utilizing both the training real images without pose labels and a CAD model, we explore a CAD-like radiance field for rendering corresponding synthetic images that have similar textures to those generated by the CAD model. Then, a backbone pose estimator trained on the synthetic data is employed to provide initial pose estimations for the synthetic images rendered from the CAD-like radiance field, and the initial object poses are refined by a global object pose refiner to generate pseudo object pose labels. Finally, the backbone pose estimator is further self-supervised as the final pose estimator by jointly utilizing the real images with pseudo object pose labels and the synthetic images rendered from the CAD-like radiance field. Experimental results on three public datasets demonstrate that ONDA-Pose significantly outperforms the comparative state-of-the-art methods in most cases.

****

Teller: Real-Time Streaming Audio-Driven Portrait Animation with Autoregressive Motion Generation

Dingcheng Zhen, Shunshun Yin, Shiyang Qin, Hou Yi, Ziwei Zhang, Siyuan Liu, Gan Qi, Ming Tao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21075-21085

In this work, we introduce the first autoregressive framework for real-time, audio-driven portrait animation, a.k.a, talking head. Beyond the challenge of lengthy animation times, a critical challenge in realistic talking head generation lies in preserving the natural movement of diverse body parts. To this end, we propose Teller, the first streaming audio-driven protrait animation framework with autoregressive motion generation. Specifically, Teller first decomposes facial and body detail animation into two components: Facial Motion Latent Generation (FMLG) based on an autoregressive transfromer, and movement authenticity refinement using a Efficient Temporal Module (ETM).Concretely, FMLG employs a Residual VQ model to map the facial motion latent from the implicit keypoint-based model into discrete motion tokens, which are then temporally sliced with audio embeddings. This enables the AR tranformer to learn real-time, stream-based mappings from audio to motion.Furthermore, Teller incorporate ETM to capture finer motion details. This module ensures the physical consistency of body parts and accessories, such as neck muscles and earrings, improving the realism of these movements.Teller is designed to be efficient, surpassing the inference speed of diffusion-based models (Hallo 20.93s vs. Teller 0.92s for one second video generation), and achieves a real-time streaming performance of up to 25 FPS. Extensive experiments demonstrate that our method outperforms recent audio-driven portrait animation models, especially in small movements, as validated by human evaluations with a significant margin in quality and realism.

****

GASP: Gaussian Avatars with Synthetic Priors

Jack Saunders, Charlie Hewitt, Yanan Jian, Marek Kowalski, Tadas Baltrusaitis, Yiye Chen, Darren Cosker, Virginia Estellers, Nicholas Gydé, Vinay P. Namboodiri, Benjamin E. Lundell; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 271-280

Gaussian Splatting has changed the game for real-time photo-realistic rendering. One of the most popular applications of Gaussian Splatting is to create animatable avatars, known as Gaussian Avatars. Recent works have pushed the boundaries of quality and rendering efficiency but suffer from two main limitations. Either they require expensive multi-camera rigs to produce avatars with free-view rendering, or they can be trained with a single camera but only rendered at high quality from this fixed viewpoint. An ideal model would be trained using a short monocular video or image from available hardware, such as a webcam, and rendered f

rom any view. To this end, we propose GASP: Gaussian Avatars with Synthetic Priors. To overcome the limitations of existing datasets, we exploit the pixel-perfect nature of synthetic data to train a Gaussian Avatar prior. By fitting this prior model to a single photo or videoand fine-tuning it, we get a high-quality Gaussian Avatar, which supports 360^\circ rendering. Our prior is only required for fitting, not inference, enabling real-time application. Through our method, we obtain high-quality, animatable Avatars from limited data which can be animated and rendered at 70fps on commercial hardware.
****

Q-PART: Quasi-Periodic Adaptive Regression with Test-time Training for Pediatric Left Ventricular Ejection Fraction Regression

Jie Liu, Tiexin Qin, Hui Liu, Yilei Shi, Lichao Mou, Xiao Xiang Zhu, Shiqi Wang, Haoliang Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15560-15569

In this work, we address the challenge of adaptive pediatric Left Ventricular Ejection Fraction (LVEF) assessment. While Test-time Training (TTT) approaches show promise for this task, they suffer from two significant limitations. Existing TTT works are primarily designed for classification tasks rather than continuous value regression, and they lack mechanisms to handle the quasi-periodic nature of cardiac signals. To tackle these issues, we propose a novel Quasi-Periodic Adaptive Regression with Test-time Training (Q-PART) framework. In the training stage, the proposed Quasi-Period Network decomposes the echocardiogram into periodic and aperiodic components within latent space by combining parameterized helix trajectories with Neural Controlled Differential Equations. During inference, our framework further employs a variance minimization strategy across image augmentations that simulate common quality issues in echocardiogram acquisition, along with differential adaptation rates for periodic and aperiodic components. Theoretical analysis is provided to demonstrate that our variance minimization objective effectively bounds the regression error under mild conditions. Furthermore, extensive experiments across three pediatric age groups demonstrate that Q-PART not only significantly outperforms existing approaches in pediatric LVEF prediction, but also exhibits strong clinical screening capability with high mAUROC scores (up to 0.9747) and maintains gender-fair performance across all metrics, validating its robustness and practical utility in pediatric echocardiography analysis.
****

Composing Parts for Expressive Object Generation

Harsh Rangwani, Aishwarya Agarwal, Kuldeep Kulkarni, R. Venkatesh Babu, Srikrishna Karanam; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13209-13219

Image composition and generation are processes where the artists need control over various parts of the generated images. However, the current state-of-the-art generation models, like Stable Diffusion, cannot handle fine-grained part-level attributes in the text prompts. Specifically, when additional attribute details are added to the base text prompt, these text-to-image models either generate an image vastly different from the image generated from the base prompt or ignore the attribute details. To mitigate these issues, we introduce PartComposer, a training-free method that enables image generation based on fine-grained part-level attributes specified for objects in the base text prompt. This allows more control for artists and enables novel object compositions by combining distinctive object parts. PartComposer first localizes object parts by denoising the object region from a specific diffusion process. This enables each part token to be localized to the right region. After obtaining part masks, we run a localized diffusion process in each part region based on fine-grained part attributes and combine them to produce the final image. All stages of PartComposer are based on repurposing a pre-trained diffusion model, which enables it to generalize across domains. We demonstrate the effectiveness of part-level control provided by PartComposer through qualitative visual examples and quantitative comparisons with contemporary baselines.
****

CarPlanner: Consistent Auto-regressive Trajectory Planning for Large-Scale Reinforcement Learning in Autonomous Driving

Dongkun Zhang, Jiaming Liang, Ke Guo, Sha Lu, Qi Wang, Rong Xiong, Zhenwei Miao, Yue Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17239-17248

Trajectory planning is vital for autonomous driving, ensuring safe and efficient navigation in complex environments. While recent learning-based methods, particularly reinforcement learning (RL), have shown promise in specific scenarios, RL planners struggle with training inefficiencies and managing large-scale, real-world driving scenarios.In this paper, we introduce CarPlanner, a Consistent auto-regressive Planner that uses RL to generate multi-modal trajectories. The auto-regressive structure enables efficient large-scale RL training, while the incorporation of consistency ensures stable policy learning by maintaining coherent temporal consistency across time steps. Moreover, CarPlanner employs a generation-selection framework with an expert-guided reward function and an invariant-view module, simplifying RL training and enhancing policy performance.Extensive analysis demonstrates that our proposed RL framework effectively addresses the challenges of training efficiency and performance enhancement, positioning CarPlanner as a promising solution for trajectory planning in autonomous driving.To the best of our knowledge, we are the first to demonstrate that the RL-based planner can surpass both IL- and rule-based state-of-the-arts (SOTAs) on the challenging large-scale real-world dataset nuPlan. Our proposed CarPlanner surpasses RL-, IL-, and rule-based SOTA approaches within this demanding dataset.

****

Apply Hierarchical-Chain-of-Generation to Complex Attributes Text-to-3D Generation

Yiming Qin, Zhu Xu, Yang Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18521-18530

Recent text-to-3D generation models have demonstrated remarkable abilities in producing high-quality 3D assets. Despite their great advancements, current models struggle to generate satisfying 3D objects with complex attributes. The difficulty for such complex attributes 3D generation arises from two aspects: (1) existing text-to-3D approaches typically lift text-to-image models to extract semantics via text encoders, while the text encoder exhibits limited comprehension ability for long descriptions, leading to deviated cross-attention focus, subsequently wrong attribute binding in generated results. (2) Objects with complex attributes often exhibit occlusion relationships between different parts, which demands a reasonable generation order as well as explicit disentanglement of different parts to enable structural coherent and attribute following results. Though some works introduce manual efforts to alleviate the above issues, their quality is unstable and highly reliant on manual information. To tackle above problems, we propose a automated method Hierarchical-Chain-of-Generation (HCoG). It leverages a large language model to analyze the long description, decomposes it into several blocks representing different object parts, and organizes an optimal generation order from in to out according to the occlusion relationship between parts, turning the whole generation process into a hierarchical chain. For optimization within each block, we first generate the necessary components coarsely, then bind their attributes precisely by target region localization and corresponding 3D Gaussian kernel optimization. For optimization between blocks, we introduce Gaussian Extension and Label Elimination to seamlessly generate new parts by extending new Gaussian kernels, re-assigning semantic labels, and eliminating unnecessary kernels, ensuring that only relevant parts are added without disrupting previously optimized parts. Experiments validate HCoG's effectiveness in handling complex attributes 3D assets and witnesses high-quality results. The code is available at https://github.com/Wakals/GASCOL.

****

PersonaHOI: Effortlessly Improving Face Personalization in Human-Object Interaction Generation

Xinting Hu, Haoran Wang, Jan Eric Lenssen, Bernt Schiele; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23775-23784

We introduce PersonaHOI, a training- and tuning-free framework that fuses a general StableDiffusion model with a personalized face diffusion (PFD) model to generate identity-consistent human-object interaction (HOI) images. While existing PFD models have advanced significantly, they often overemphasize facial features at the expense of full-body coherence, PersonaHOI introduces an additional StableDiffusion (SD) branch guided by HOI-oriented text inputs. By incorporating cross-attention constraints in the PFD branch and spatial merging at both latent and residual levels, PersonaHOI preserves personalized facial details while ensuring interactive non-facial regions. Experiments, validated by a novel interaction alignment metric, demonstrate the superior realism and scalability of PersonaHOI, establishing a new standard for practical personalized face with HOI generation. Code is available at https://github.com/JoyHuYY1412/PersonaHOI.
****

Video-Panda: Parameter-efficient Alignment for Encoder-free Video-Language Models

Jinhui Yi, Syed Talal Wasim, Yanan Luo, Muzammal Naseer, Juergen Gall; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24119-24128

We present an efficient encoder-free approach for video-language understanding that achieves competitive performance while significantly reducing computational overhead. Current video-language models typically rely on heavyweight image encoders (300M-1.1B parameters) or video encoders (1B-1.4B parameters), creating a substantial computational burden when processing multi-frame videos. Our method introduces a novel Spatio-Temporal Alignment Block (STAB) that directly processes video inputs without requiring pre-trained encoders while using only 45M parameters for visual processing - at least a 6.5x reduction compared to traditional approaches. The STAB architecture combines Local Spatio-Temporal Encoding for fine-grained feature extraction, efficient spatial downsampling through learned attention and separate mechanisms for modeling frame-level and video-level relationships. Our model achieves comparable or superior performance to encoder-based approaches for open-ended video question answering on standard benchmarks. The fine-grained video question-answering evaluation demonstrates our model's effectiveness, outperforming the encoder-based approaches Video-ChatGPT and Video-LLaVA in key aspects like correctness and temporal understanding. Extensive ablation studies validate our architectural choices and demonstrate the effectiveness of our spatio-temporal modeling approach while achieving 3-4x faster processing speeds than previous methods. Code is available at https://jh-yi.github.io/Video-Panda.
****

COSMIC: Clique-Oriented Semantic Multi-space Integration for Robust CLIP Test-Time Adaptation

Fanding Huang, Jingyan Jiang, Qinting Jiang, Hebei Li, Faisal Nadeem Khan, Zhi Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9772-9781

Recent vision-language models (VLMs) face significant challenges in test-time adaptation to novel domains. While cache-based methods show promise by leveraging historical information, they struggle with both caching unreliable feature-label pairs and indiscriminately using single-class information during querying, significantly compromising adaptation accuracy. To address these limitations, we propose COSMIC (\underline C lique-\underline O riented \underline S emantic \underline M ulti-space \underline I ntegration for \underline C LIP), a robust test-time adaptation framework that enhances adaptability through multi-granular, cross-modal semantic caching and graph-based querying mechanisms. Our framework introduces two key innovations: Dual Semantics Graph (DSG) and Clique Guided Hyper-class (CGH). The Dual Semantics Graph constructs complementary semantic spaces by incorporating textual features, coarse-grained CLIP features, and fine-grained DINOv2 features to capture rich semantic relationships. Building upon these dual graphs, the Clique Guided Hyper-class component leverages structured class relationships to enhance prediction robustness through correlated class selection. Extensive experiments demonstrate COSMIC's superior performance across multiple b

enchmarks, achieving significant improvements over state-of-the-art methods: 15.81% gain on out-of-distribution tasks and 5.33% on cross-domain generation with CLIP RN-50.
****

MonoSplat: Generalizable 3D Gaussian Splatting from Monocular Depth Foundation Models
Yifan Liu, Keyu Fan, Weihao Yu, Chenxin Li, Hao Lu, Yixuan Yuan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21570-21579
Recent advances in generalizable 3D Gaussian Splatting have demonstrated promising results in real-time high-fidelity rendering without per-scene optimization, yet existing approaches still struggle to handle unfamiliar visual content during inference on novel scenes due to limited generalizability. To address this challenge, we introduce MonoSplat, a novel framework that leverages rich visual priors from pre-trained monocular depth foundation models for robust Gaussian reconstruction. Our approach consists of two key components: a Mono-Multi Feature Adapter that transforms monocular features into multi-view representations, coupled with an Integrated Gaussian Prediction module that effectively fuses both feature types for precise Gaussian generation. Through the Adapter's lightweight attention mechanism, features are seamlessly aligned and aggregated across views while preserving valuable monocular priors, enabling the Prediction module to generate Gaussian primitives with accurate geometry and appearance. Through extensive experiments on diverse real-world datasets, we convincingly demonstrate that MonoSplat achieves superior reconstruction quality and generalization capability compared to existing methods while maintaining computational efficiency with minimal trainable parameters. Codes are available at \href https://github.com/CUHK-AIM-Group/MonoSplat https://github.com/CUHK-AIM-Group/MonoSplat .
****

Hybrid Global-Local Representation with Augmented Spatial Guidance for Zero-Shot Referring Image Segmentation
Ting Liu, Siyuan Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29634-29643
Recent advances in zero-shot referring image segmentation (RIS), driven by models such as the Segment Anything Model (SAM) and CLIP, have made substantial progress in aligning visual and textual information. Despite these successes, the extraction of precise and high-quality mask region representations remains a critical challenge, limiting the full potential of RIS tasks. In this paper, we introduce a training-free, hybrid global-local feature extraction approach that integrates detailed mask-specific features with contextual information from the surrounding area, enhancing mask region representation. To further strengthen alignment between mask regions and referring expressions, we propose a spatial guidance augmentation strategy that improves spatial coherence, which is essential for accurately localizing described areas. By incorporating multiple spatial cues, this approach facilitates more robust and precise referring segmentation. Extensive experiments on standard RIS benchmarks demonstrate that our method significantly outperforms existing zero-shot referring segmentation models, achieving substantial performance gains. We believe our approach advances RIS tasks and establishes a versatile framework for region-text alignment, offering broader implications for cross-modal understanding and interaction. The code will be publicly available.
****

SOLVE: Synergy of Language-Vision and End-to-End Networks for Autonomous Driving
Xuesong Chen, Linjiang Huang, Tao Ma, Rongyao Fang, Shaoshuai Shi, Hongsheng Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12068-12077
The integration of Vision-Language Models (VLMs) into autonomous driving systems has shown promise in addressing key challenges such as learning complexity, interpretability, and common-sense reasoning. However, existing approaches often struggle with efficient integration and real-time decision-making due to computational demands. In this paper, we introduce SOLVE, an innovative framework that sy

nergizes VLMs with end-to-end (E2E) models to enhance autonomous vehicle plannin
g. Our approach emphasizes knowledge sharing at the feature level through a shar
ed visual encoder, enabling comprehensive interaction between VLM and E2E compon
ents. We propose a Trajectory Chain-of-Thought (T-CoT) paradigm, which progressi
vely refines trajectory predictions, reducing uncertainty and improving accuracy
. By employing a temporal decoupling strategy, SOLVE achieves efficient asynchro
nous cooperation, aligning high-quality VLM outputs with E2E real-time performan
ce. Evaluated on the nuScenes dataset, our method demonstrates significant impro
vements in trajectory prediction accuracy, paving the way for more robust and re
liable autonomous driving systems.
****

Shift the Lens: Environment-Aware Unsupervised Camouflaged Object Detection
Ji Du, Fangwei Hao, Mingyang Yu, Desheng Kong, Jiesheng Wu, Bin Wang, Jing Xu, P
ing Li; Proceedings of the Computer Vision and Pattern Recognition Conference (C
VPR), 2025, pp. 19271-19282
Camouflaged Object Detection (COD) seeks to distinguish objects from their highl
y similar backgrounds. Existing work has essentially focused on isolating camouf
laged objects from the environment, demonstrating ever-improving performance but
 at the cost of extensive annotations and complex optimizations. In this paper,
we diverge from this paradigm and shift the lens to isolating the salient enviro
nment from the camouflaged object. We introduce EASE, an Environment-Aware unSup
Ervised COD framework that identifies the environment by referencing an environm
ent prototype library and detects camouflaged objects by inverting the retrieved
 environmental features. Specifically, our approach (DiffPro) uses large multimo
dal models, diffusion models, and vision-foundation models to construct the envi
ronment prototype library. To retrieve environments from the library and refrain
 from confusing foreground and background, we incorporate three retrieval scheme
s: Kernel Density Estimation-based Adaptive Threshold (KDE-AT), Global-to-Local
pixel-level retrieval (G2L), and Self-Retrieval (SR). Our experiments demonstrat
e significant improvements over current unsupervised methods, with EASE achievin
g an average gain of over 10% on the COD10K dataset. When integrated with SAM, E
ASE surpasses prompt-based segmentation approaches and performs competitively wi
th state-of-the-art fully-supervised methods. Code is available at https://githu
b.com/xiaohainku/EASE.
****

Probability Density Geodesics in Image Diffusion Latent Space
Qingtao Yu, Jaskirat Singh, Zhaoyuan Yang, Peter Henry Tu, Jing Zhang, Hongdong
Li, Richard Hartley, Dylan Campbell; Proceedings of the Computer Vision and Patt
ern Recognition Conference (CVPR), 2025, pp. 27989-27998
Diffusion models indirectly estimate the probability density over a data space,
which can be used to study its structure. In this work, we show that geodesics c
an be computed in diffusion latent space, where the norm induced by the spatiall
y-varying inner product is inversely proportional to the probability density. In
 this formulation, a path that traverses a high density (that is, probable) regi
on of image latent space is shorter than the equivalent path through a low densi
ty region. We present algorithms for solving the associated initial and boundary
 value problems and show how to compute the probability density along the path a
nd the geodesic distance between two points. Using these techniques, we analyze
how closely video clips approximate geodesics in a pre-trained image diffusion s
pace. Finally, we demonstrate how these techniques can be applied to training-fr
ee image sequence interpolation and extrapolation, given a pre-trained image dif
fusion model.
****

High-quality Point Cloud Oriented Normal Estimation via Hybrid Angular and Eucli
dean Distance Encoding
Yuanqi Li, Jingcheng Huang, Hongshen Wang, Peiyuan Lv, Yansong Liu, Jiuming Zhen
g, Jie Guo, Yanwen Guo; Proceedings of the Computer Vision and Pattern Recogniti
on Conference (CVPR), 2025, pp. 1287-1296
The proliferation of Light Detection and Ranging (LiDAR) technology has facilita
ted the acquisition of three-dimensional point clouds, which are integral to app

lications in VR, AR, and Digital Twin. Oriented normals, critical for 3D reconst ruction and scene analysis, cannot be directly extracted from scenes using LiDAR due to its operational principles. Previous traditional or learning-based metho ds are prone to inaccuracies due to uneven distribution and noise due to the dep endence on local geometry features. This paper addresses the challenge of estima ting oriented point normals by introducing a point cloud normal estimation frame work via hybrid angular and Euclidean distance encoding (HAE). Our method overco mes the limitations of local geometric information by combining angular and Eucl idean spaces to extract features from both point cloud coordinates and light ray s, leading to more accurate normal estimation. The core of our network consists of an angular distance encoding module, which leverages both ray directions and point coordinates for unoriented normal refinement, and a ray feature fusion mod ule for normal orientation, that is robust to noise. We also provide a point clo ud dataset with ground truth normals, generated a virtual scanner, which reflect s real scanning distributions and noise profiles.

****

DriveScape: High-Resolution Driving Video Generation by Multi-View Feature Fusio n

Wei Wu, Xi Guo, Weixuan Tang, Tingxuan Huang, Chiyu Wang, Chenjing Ding; Proceed ings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17187-17196

Recent advancements in generative models offer promising solutions for synthesiz ing realistic driving videos, aiding in training autonomous driving perception m odels. However, existing methods often struggle with high-resolution multi-view generation, mainly due to the significant memory and computational overhead caus ed by simultaneously inputting multi-view videos into denoising diffusion models .In this paper, we propose a driving video generation framework based on multi-v iew feature fusion named DriveScape for multi-view 3D condition-guided video gen eration. We introduce a Bi-Directional Modulated Transformer (BiMoT) module to e ncode, fuse and inject multi-view features along with various 3D road structures and objects, which enables high-resolution multi-view generation. Consequently, our approach allows precise control over video generation, greatly enhancing re alism and providing a robust solution for creating high-quality, multi-view driv ing videos.Our framework achieves state-of-the-art results on the nuScenes datas et, demonstrating impressive generative quality metrics with an FID score of 8.3 4 and an FVD score of 76.39, as well as superior performance across various perc eption tasks. This lays the foundation for more accurate environment simulation in autonomous driving. We plan to make our code and pre-trained model publicly a vailable.Please refer to index.html webpage in the supplementary materials for m ore visualization results.

****

Training-free Neural Architecture Search through Variance of Knowledge of Deep N etwork Weights

Ondrej Tybl, Lukas Neumann; Proceedings of the Computer Vision and Pattern Recog nition Conference (CVPR), 2025, pp. 14881-14890

Deep learning has revolutionized computer vision, but it achieved its tremendous success using deep network architectures which are mostly hand-crafted and ther efore likely suboptimal. Neural Architecture Search (NAS) aims to bridge this ga p by following a well-defined optimization paradigm which systematically looks f or the best architecture, given objective criterion such as maximal classificati on accuracy. The main limitation of NAS is however its astronomical computationa l cost, as it typically requires training each candidate network architecture fr om scratch.In this paper, we aim to alleviate this limitation by proposing a nov el training-free proxy for image classification accuracy based on Fisher Informa tion. The proposed proxy has a strong theoretical background in statistics and i t allows estimating expected image classification accuracy of a given deep netwo rk without training the network, thus significantly reducing computational cost of standard NAS algorithms. Our training-free proxy achieves state-of-the-art re sults on three public datasets and in two search spaces, both when evaluated usi ng previously proposed metrics, as well as using a new metric that we propose wh

ich we demonstrate is more informative for practical NAS applications. The sourc
e code is publicly available.
****

Every SAM Drop Counts: Embracing Semantic Priors for Multi-Modality Image Fusion
 and Beyond
Guanyao Wu, Haoyu Liu, Hongming Fu, Yichuan Peng, Jinyuan Liu, Xin Fan, Risheng
Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR
), 2025, pp. 17882-17891
Multi-modality image fusion, particularly infrared and visible, plays a crucial
role in integrating diverse modalities to enhance scene understanding. Although
early research prioritized visual quality, preserving fine details and adapting
to downstream tasks remains challenging. Recent approaches attempt task-specific
 design but rarely achieve "The Best of Both Worlds" due to inconsistent optimiz
ation goals. To address these issues, we propose a novel method that leverages t
he semantic knowledge from the Segment Anything Model (SAM) to grow the quality
of fusion results and enable downstream task adaptability, namely SAGE. Specific
ally, we design a Semantic Persistent Attention (SPA) Module that efficiently ma
intains source information via the persistent repository while extracting high-l
evel semantic priors from SAM. More importantly, to eliminate the impractical de
pendence on SAM during inference, we introduce a bi-level optimization-driven di
stillation mechanism with triplet losses, which allow the student network to eff
ectively extract knowledge. Extensive experiments show that our method achieves
a balance between high-quality visual results and downstream task adaptability w
hile maintaining practical deployment efficiency. The code is available at https
://github.com/RollingPlain/SAGE_IVIF.
****

EgoLife: Towards Egocentric Life Assistant
Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zh
ang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin,
 Marco Cominelli, Zhongang Cai, Bo Li, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Ho
ng, Joerg Widmer, Francesco Gringoli, Lei Yang, Ziwei Liu; Proceedings of the Co
mputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28885-28900
We introduce EgoLife, a project to develop an egocentric life assistant that acc
ompanies and enhances personal efficiency through AI-powered wearable glasses. T
o lay the foundation for this assistant, we conducted a comprehensive data colle
ction study where six participants lived together for one week, continuously rec
ording their daily activities - including discussions, shopping, cooking, social
izing, and entertainment - using AI glasses for multimodal egocentric video capt
ure, along with synchronized third-person-view video references. This effort res
ulted in the EgoLife Dataset, a comprehensive 300-hour egocentric, interpersonal
, multiview, and multimodal daily life dataset with intensive annotation. Levera
ging this dataset, we introduce EgoLifeQA, a suite of long-context, life-oriente
d question-answering tasks designed to provide meaningful assistance in daily li
fe by addressing practical questions such as recalling past relevant events, mon
itoring health habits, and offering personalized recommendations. To address the
 key technical challenges of (1) developing robust visual-audio models for egoce
ntric data, (2) enabling identity recognition, and (3) facilitating long-context
 question answering over extensive temporal information, we introduce EgoButler,
 an integrated system comprising EgoGPT and EgoRAG. EgoGPT is an omni-modal mode
l trained on egocentric datasets, achieving state-of-the-art performance on egoc
entric video understanding. EgoRAG is a retrieval-based component that supports
answering ultra-long-context questions. Our experimental studies verify their wo
rking mechanisms and reveal critical factors and bottlenecks, guiding future imp
rovements. By releasing our datasets, models, and benchmarks, we aim to stimulat
e further research in egocentric AI assistants.
****

RAEncoder: A Label-Free Reversible Adversarial Examples Encoder for Dataset Inte
llectual Property Protection
Fan Xing, Zhuo Tian, Xuefeng Fan, Xiaoyi Zhou; Proceedings of the Computer Visio
n and Pattern Recognition Conference (CVPR), 2025, pp. 20665-20674

Reversible Adversarial Examples (RAE) are designed to protect the intellectual p
roperty of datasets. Such examples can function as imperceptible adversarial exa
mples to erode the model performance of unauthorized users while allowing author
ized users to remove the adversarial perturbations and recover the original samp
les for normal model training. With the rise of Self-Supervised Learning (SSL),
an increasing number of unlabeled datasets and pre-trained encoders are availabl
e in the community. However, existing RAE methods not only rely on well-labeled
datasets for training Supervised Learning (SL) models but also exhibit poor adve
rsarial transferability when attacking SSL pre-trained encoders. To address thes
e challenges, we propose RAEncoder, the first framework for RAEs without the nee
d for labeled samples. RAEncoder aims to generate universal adversarial perturba
tions by targeting SSL pre-trained encoders. Unlike traditional RAE approaches,
the pre-trained encoder outputs the feature distribution of the protected datase
t rather than classification labels, enhancing both the attack success rate and
transferability of RAEs. Extensive experiments are conducted on six pre-trained
encoders and four SL models, covering aspects such as imperceptibility and trans
ferability. Our results demonstrate that RAEncoder effectively protects unlabele
d datasets from malicious infringements. Additional robustness experiments furth
er confirm the security of RAEncoder in practical application scenarios.
****

BrepGiff: Lightweight Generation of Complex B-rep with 3D GAT Diffusion
Hao Guo, Xiaoshui Huang, Hao jiacheng, Yunpeng Bai, Hongping Gan, Yilei Shi; Pro
ceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025,
 pp. 26587-26596
Despite advancements in Computer-Aided-Design (CAD) generation, direct generatio
n of complex Boundary Representation (B-rep) CAD models remains challenging. Thi
s difficulty arises from the parametric nature of B-rep data, complicating the e
ncoding and generation of its geometric and topological information. To address
this, we introduce BrepGiff, a lightweight generation approach for high-quality
and complex B-rep model based on 3D Graph Diffusion. First, we transfer B-rep mo
dels into 3D graphs representation. Specifically, BrepGiff extracts and integrat
es topological and geometric features to construct a 3D graph where nodes corres
pond to face centroids in 3D space, preserving adjacency and degree information.
 Geometric features are derived by sampling points in the UV domain and extracti
ng face and edge features. Then, BrepGiff applies a Graph Attention Network (GAT
) to enforce topological constraints from local to global during the degree-guid
ed diffusion process. With the 3D graph representation and efficient diffusion p
rocess, our method significantly reduces the computational cost and improves the
 quality, thus achieving lightweight generation of complex models. Experiments s
how that BrepGiff can generate complex B-rep models (>100 faces) using only 2 RT
X4090 GPUs, achieving state-of-the-art performance in B-rep generation.
****

Towards Fine-Grained Interpretability: Counterfactual Explanations for Misclassi
fication with Saliency Partition
Lintong Zhang, Kang Yin, Seong-Whan Lee; Proceedings of the Computer Vision and
Pattern Recognition Conference (CVPR), 2025, pp. 30053-30062
Attribution-based explanation techniques capture key patterns to enhance visual
interpretability. However, these patterns often lack the granularity needed for
insight in fine-grained tasks, particularly in cases of model misclassification,
 where explanations may be insufficiently detailed. To address this limitation,
we propose a fine-grained counterfactual explanation framework that generates bo
th object-level and part-level interpretability, addressing two fundamental ques
tions: (1) which fine-grained features contribute to model misclassification, an
d (2) where dominant local features influence counterfactual adjustments. Our ap
proach yields explainable counterfactuals in a non-generative manner by quantify
ing similarity and weighting component contributions within regions of interest
between correctly classified and misclassified samples. Furthermore, we introduc
e an importance-isolation module grounded in Shapley value contributions, isolat
ing features with region-specific relevance. Extensive experiments demonstrate t
he superiority of our approach in capturing more granular, intuitively meaningfu

l regions, surpassing fine-grained methods.
****

Prior-free 3D Object Tracking
Xiuqiang Song, Li Jin, Zhengxian Zhang, Jiachen Li, Fan Zhong, Guofeng Zhang, Xueying Qin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1200-1209
In this paper, we introduce a novel, truly prior-free 3D object tracking method that operates without given any model or training priors. Unlike existing methods that typically require pre-defined 3D models or specific training datasets as priors, which limit their applicability, our method is free from these constraints. Our method consists of a geometry generation module and a pose optimization module. Its core idea is to enable these two modules to automatically and iteratively enhance each other, thereby gradually building all the necessary information for the tracking task. We thus call the method as Bidirectional Iterative Tracking(BIT). The geometry generation module starts without priors and gradually generates high-precision mesh models for tracking, while the pose optimization module generates additional data during object tracking to further refine the generated models. Moreover, the generated 3D models can be stored and easily reused, allowing for seamless integration into various other tracking systems, not just our methods. Experimental results demonstrate that BIT outperforms many existing methods, even those that extensively utilize prior knowledge, while BIT does not rely on such information. Additionally, the generated 3D models deliver results comparable to actual 3D models, highlighting their superior and innovative qualities. The code is available at https://github.com/songxiuqiang/BIT.git.
****

LatentHOI: On the Generalizable Hand Object Motion Generation with Latent Hand Diffusion.
Muchen Li, Sammy Christen, Chengde Wan, Yujun Cai, Renjie Liao, Leonid Sigal, Shugao Ma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17416-17425
Current research on generating 3D hand-object interaction motion primarily focuses on in-domain objects. Generalization to unseen objects is essential for practical applications, yet it remains both challenging and largely unexplored.In this paper, we propose LatentHOI, a novel approach designed to tackle the challenges of generalizing hand-object interaction to unseen objects.Our main insight lies in decoupling high-level temporal motion from fine-grained spatial hand-object interactions with a latent diffusion model coupled with a Grasping Variational Autoencoder (GraspVAE). This configuration not only enhances the conditional dependency between spatial grasp and temporal motion but also improves data utilization and reduces overfitting through regularization in the latent space. We conducted extensive experiments in an unseen-object setting on both single-hand grasping and bi-manual motion datasets, including GRAB, DexYCB, and OakInk.Quantitative and qualitative evaluations demonstrate that our method significantly enhances the realism and physical plausibility of generated motions for unseen objects, both in single and bimanual manipulations, compared to the state-of-the-art.
****

Progressive Correspondence Regenerator for Robust 3D Registration
Guiyu Zhao, Sheng Ao, Ye Zhang, Kai Xu, Yulan Guo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1210-1219
Obtaining enough high-quality correspondences is crucial for robust registration. Existing correspondence refinement methods mostly follow the paradigm of outlier removal, which either fails to correctly identify the accurate correspondences under extreme outlier ratios, or select too few correct correspondences to support robust registration. To address this challenge, we propose a novel approach named Regor, which is a progressive correspondence regenerator that generates higher-quality matches whist sufficiently robust for numerous outliers. In each iteration, we first apply prior-guided local grouping and generalized mutual matching to generate the local region correspondences. A powerful center-aware three-point consistency is then presented to achieve local correspondence correction, instead of removal. Further, we employ global correspondence refinement to obta

in accurate correspondences from a global perspective. Through progressive itera tions, this process yields a large number of high-quality correspondences. Exten sive experiments on both indoor and outdoor datasets demonstrate that the propos ed Regor significantly outperforms existing outlier removal techniques. More cri tically, our approach obtain 10 times more correct correspondences than outlier removal methods. As a result, our method is able to achieve robust registration even with weak features. The code is available at https://github.com/GuiyuZhao/R egor.
****

Cross-Modal 3D Representation with Multi-View Images and Point Clouds
Ziyang Zhou, Pinghui Wang, Zi Liang, Haitao Bai, Ruofei Zhang; Proceedings of th e Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3728-3739
The advancement of 3D understanding and representation is a crucial step for the next phase of autonomous driving, robotics, augmented and virtual reality, 3D g aming and 3D e-commerce products. However, existing 3D semantic representation r esearch has primarily focused on point clouds to perceive 3D objects and scenes, overlooking the rich visual details offered by multi-view images, thereby limit ing the potential of 3D semantic representation. This paper introduces OpenView, a novel representation method that integrates both point clouds and multi-view images to form a unified 3D representation. OpenView comprises a unique fusion f ramework, sequence-independent modeling, a cross-modal fusion encoder, and a pro gressive hard learning strategy. Our experiments demonstrate that OpenView outpe rforms the state-of-the-art by 11.5% and 5.5% on the R@1 metric for cross-modal retrieval and the Top-1 metric for zero-shot classification tasks, respectively. Furthermore, we showcase some applications of OpenView: 3D retrieval, 3D captio ning and hierarchical data clustering, highlighting its generality in the field of 3D representation learning.
****

Chapter-Llama: Efficient Chaptering in Hour-Long Videos with LLMs
Lucas Ventura, Antoine Yang, Cordelia Schmid, Gül Varol; Proceedings of the Comp uter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18947-18958
We address the task of video chaptering, i.e., partitioning a long video timelin e into semantic units and generating corresponding chapter titles. While relativ ely underexplored, automatic chaptering has the potential to enable efficient na vigation and content retrieval in long-form videos. In this paper, we achieve st rong chaptering performance on hour-long videos by efficiently addressing the pr oblem in the text domain with our "Chapter-Llama" framework. Specifically, we le verage a pre-trained large language model (LLM) with large context window, and f eed as input (i) speech transcripts and (ii) captions describing video frames, a long with their respective timestamps. Given the inefficiency of exhaustively ca ptioning all frames, we propose a lightweight speech-guided frame selection stra tegy based on speech transcript content, and experimentally demonstrate remarkab le advantages. We train the LLM to output timestamps for the chapter boundaries, as well as free-form chapter titles. This simple yet powerful approach scales t o processing one-hour long videos in a single forward pass. Our results demonstr ate substantial improvements (e.g., 45.3 vs 26.7 F1 score) over the state of the art on the recent VidChapters-7M benchmark. To promote further research, we rel ease our code and models at our project page.
****

Decompositional Neural Scene Reconstruction with Generative Diffusion Prior
Junfeng Ni, Yu Liu, Ruijie Lu, Zirui Zhou, Song-Chun Zhu, Yixin Chen, Siyuan Hua ng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) , 2025, pp. 6022-6033
Decompositional reconstruction of 3D scenes, with complete shapes and detailed t exture of all objects within, is intriguing for downstream applications but rema ins challenging, particularly with sparse views as input. Recent approaches inco rporate semantic or geometric regularization to address this issue, but they suf fer significant degradation in underconstrained areas and fail to recover occlud ed regions. We argue that the key to solving this problem lies in supplementing missing information for these areas. To this end, we propose DP-Recon, which emp

loys diffusion priors in the form of Score Distillation Sampling (SDS) to optimi ze the neural representation of each individual object under novel views. This p rovides additional information for the underconstrained areas, but directly inco rporating diffusion prior raises potential conflicts between the reconstruction and generative guidance. Therefore, we further introduce a visibility-guided app roach to dynamically adjust the per-pixel SDS loss weights. Together these compo nents enhance both geometry and appearance recovery while remaining faithful to input images. Extensive experiments across Replica and ScanNet++ demonstrate th at our method significantly outperforms state-of-the-art methods. Notably, it ac hieves better object reconstruction under 10 views than the baselines under 100 views. Our method enables seamless text-based editing for geometry and appearanc e through SDS optimization and produces decomposed object meshes with detailed U V maps that support photorealistic Visual effects (VFX) editing. The project pag e is available at https://dp-recon.github.io/.

****

Distribution Prototype Diffusion Learning for Open-set Supervised Anomaly Detect ion

Fuyun Wang, Tong Zhang, Yuanzhi Wang, Yide Qiu, Xin Liu, Xu Guo, Zhen Cui; Proce edings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, p p. 20416-20426

In Open-set Supervised Anomaly Detection (OSAD), the existing methods typically generate pseudo anomalies to compensate for the scarcity of observed anomaly s amples, while overlooking critical priors of normal samples, leading to less e ffective discriminative boundaries. To address this issue, we propose a Distri bution Prototype Diffusion Learning (DPDL) method aimed at enclosing normal samp les within a compact and discriminative distribution space. Specifically, we c onstruct multiple learnable Gaussian prototypes to create a latent representati on space for abundant and diverse normal samples and learn a Schr"odinger brid ge to facilitate a diffusive transition toward these prototypes for normal samp les while steering anomaly samples away. Moreover, to enhance inter-sample sep aration, we design a dispersion feature learning way in hyperspherical space, which benefits the identification of out-ofdistribution anomalies. Experimenta l results demonstrate the effectiveness and superiority of our proposed DPDL, achieving state-of-the-art performance on 9 public datasets.

****

Learning Visual Generative Priors without Text

Shuailei Ma, Kecheng Zheng, Ying Wei, Wei Wu, Fan Lu, Yifei Zhang, Chen-Wei Xie, Biao Gong, Jiapeng Zhu, Yujun Shen; Proceedings of the Computer Vision and Patt ern Recognition Conference (CVPR), 2025, pp. 8051-8061

Although text-to-image (T2I) models have recently thrived as visual generative p riors, their reliance on high-quality text-image pairs makes scaling up expensiv e. We argue that grasping the cross-modality alignment is not a necessity for a sound visual generative prior, whose focus should be on texture modeling. Such a philosophy inspires us to study image-to-image (I2I) generation, where models c an learn from in-the-wild images in a self-supervised manner. We first develop a pure vision-based training framework, Lumos, and confirm the feasibility and th e scalability of learning I2I models. We then find that, as an upstream task of T2I, our I2I model serves as a more foundational visual prior and achieves on-pa r or better performance than existing T2I models using only 1/10 text-image pair s for fine-tuning. We further demonstrate the superiority of I2I priors over T2I priors on some text-irrelevant vision tasks, like image-to-3D and image-to-vide o.

****

Joint Scheduling of Causal Prompts and Tasks for Multi-Task Learning

Chaoyang Li, Jianyang Qin, Jinhao Cui, Zeyu Liu, Ning Hu, Qing Liao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 251 24-25134

Multi-task prompt learning has emerged as a promising technique for fine-tuning pre-trained Vision-Language Models (VLMs) to various downstream tasks. However, existing methods ignore challenges caused by spurious correlations and dynamic t

ask relationships, which may reduce the model performance. To tackle these chall enges, we propose JSCPT, a novel approach for Joint Scheduling of Causal Prompts and Tasks to enhance multi-task prompt learning. Specifically, we first design a Multi-Task Vison-Language Prompt (MTVLP) model, which learns task-shared and t ask-specific vison-language prompts and selects useful prompt features via causa l intervention, alleviating spurious correlations. Then, we propose the task-pro mpt scheduler that models inter-task affinities and assesses the causal effect o f prompt features to optimize the multi-task prompt learning process. Finally, w e formulate the scheduler and the multi-task prompt learning process as a bi-lev el optimization problem to optimize prompts and tasks adaptively. In the lower o ptimization, MTVLP is updated with the scheduled gradient, while in the upper op timization, the scheduler is updated with the implicit gradient. Extensive exper iments show the superiority of our proposed JSCPT approach over several baseline s in terms of multi-task prompt learning for pre-trained VLMs.

****

Full-DoF Egomotion Estimation for Event Cameras Using Geometric Solvers

Ji Zhao, Banglei Guan, Zibin Liu, Laurent Kneip; Proceedings of the Computer Vis ion and Pattern Recognition Conference (CVPR), 2025, pp. 11515-11524

For event cameras, current sparse geometric solvers for egomotion estimation ass ume that the rotational displacements are known, such as those provided by an IM U. Thus, they can only recover the translational motion parameters. Recovering f ull-DoF motion parameters using a sparse geometric solver is a more challenging task, and has not yet been investigated. In this paper, we propose several solve rs to estimate both rotational and translational velocities within a unified fra mework. Our method leverages event manifolds induced by line segments. The probl em formulations are based on either an incidence relation for lines or a novel c oplanarity relation for normal vectors. We demonstrate the possibility of recove ring full-DoF egomotion parameters for both angular and linear velocities withou t requiring extra sensor measurements or motion priors. To achieve efficient op timization, we exploit the Adam framework with a first-order approximation of ro tations for quick initialization. Experiments on both synthetic and real-world d ata demonstrate the effectiveness of our method. The code is available at https: //github.com/jizhaox/relpose-event.

****

3DGUT: Enabling Distorted Cameras and Secondary Rays in Gaussian Splatting

Qi Wu, Janick Martinez Esturo, Ashkan Mirzaei, Nicolas Moënne-Loccoz, Zan Gojcic ; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26036-26046

3D Gaussian Splatting (3DGS) enables efficient reconstruction and high-fidelity real-time rendering of complex scenes on consumer hardware. However, due to its rasterization-based formulation, 3DGS is constrained to ideal pinhole cameras a nd lacks support for secondary lighting effects. Recent methods address these li mitations by tracing the particles instead, but, this comes at the cost of signi ficantly slower rendering. In this work, we propose 3D Gaussian Unscented Transf orm (3DGUT), replacing the EWA splatting formulation with the Unscented Transfor m that approximates the particles through sigma points, which can be projected e xactly under any nonlinear projection function. This modification enables trivia l support of distorted cameras with time dependent effects such as rolling shutt er, while retaining the efficiency of rasterization. Additionally, we align our rendering formulation with that of tracing-based methods, enabling secondary ray tracing required to represent phenomena such as reflections and refraction with in the same 3D representation. The source code is available at: https://github.c om/nv-tlabs/3dgrut.

****

Teaching Large Language Models to Regress Accurate Image Quality Scores Using Sc ore Distribution

Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, Chao Dong; Proceedings of the Comp uter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14483-14494

With the rapid advancement of Multi-modal Large Language Models (MLLMs), MLLM-ba sed Image Quality Assessment (IQA) methods have shown promising performance in l

inguistic quality description. However, current methods still fall short in accu rately scoring image quality. In this work, we aim to leverage MLLMs to regress accurate quality scores. A key challenge is that the quality score is inherently continuous, typically modeled as a Gaussian distribution, whereas MLLMs generat e discrete token outputs. This mismatch necessitates score discretization. Previ ous approaches discretize the mean score into a one-hot label, resulting in info rmation loss and failing to capture inter-image relationships. We propose a dist ribution-based approach that discretizes the score distribution into a soft labe l. This method preserves the characteristics of the score distribution, achievin g high accuracy and maintaining inter-image relationships. Moreover, to address dataset variation, where different IQA datasets exhibit various distributions, w e introduce a fidelity loss based on Thurstone's model. This loss captures intra -dataset relationships, facilitating co-training across multiple IQA datasets. W ith these designs, we develop the distribution-based Depicted image Quality Asse ssment model for Score regression (DeQA-Score). Experiments across multiple benc hmarks show that DeQA-Score stably outperforms baselines in score regression. Al so, DeQA-Score can predict the score distribution that closely aligns with human annotations. Codes and model weights have been released in https://depictqa.git hub.io/deqa-score/.

****

Lifting the Veil on Visual Information Flow in MLLMs: Unlocking Pathways to Fast er Inference

Hao Yin, Guangzong Si, Zilei Wang; Proceedings of the Computer Vision and Patter n Recognition Conference (CVPR), 2025, pp. 9382-9391

Multimodal large language models (MLLMs) improve performance on vision-language tasks by integrating visual features from pre-trained vision encoders into large language models (LLMs). However, how MLLMs process and utilize visual informati on remains unclear. In this paper, a shift in the dominant flow of visual inform ation is uncovered: (1) in shallow layers, strong interactions are observed betw een image tokens and instruction tokens, where most visual information is inject ed into instruction tokens to form cross-modal semantic representations; (2) in deeper layers, image tokens primarily interact with each other, aggregating the remaining visual information to optimize semantic representations within the vis ual modality. Based on these insights, we propose Hierarchical Modality-Aware Pr uning (HiMAP), a plug-and-play inference acceleration method that dynamically pr unes image tokens at specific layers, reducing computational costs by approximat ely 65% without sacrificing performance. Our findings offer a new understanding of visual information processing in MLLMs and provide a state-of-the-art solutio n for efficient inference.

****

Your Scale Factors are My Weapon: Targeted Bit-Flip Attacks on Vision Transforme rs via Scale Factor Manipulation

Jialai Wang, Yuxiao Wu, Weiye Xu, Yating Huang, Chao Zhang, Zongpeng Li, Mingwei Xu, Zhenkai Liang; Proceedings of the Computer Vision and Pattern Recognition C onference (CVPR), 2025, pp. 20103-20112

Vision Transformers (ViTs) have experienced significant progress and are quantiz ed for deployment in resource-constrained applications. Quantized models are vul nerable to targeted bit-flip attacks (BFAs). A targeted BFA prepares a trigger a nd a corresponding Trojan/backdoor, inserting the latter (with RowHammer bit fli pping) into a victim model, to mislead its classification on samples containing the trigger. Existing targeted BFAs on quantized ViTs are limited in that: (1) t hey require numerous bit-flips, and (2) the separation between flipped bits is b elow 4 KB, making attacks infeasible with RowHammer in real-world scenarios. We propose a new and practical targeted attack Flip-S against quantized ViTs. The c ore insight is that in quantized models, a scale factor change ripples through a batch of model weights. Consequently, flipping bits in scale factors, rather th an solely in model weights, enables more cost-effective attacks. We design a Sca le-Factor-Search (SFS) algorithm to identify critical bits in scale factors for flipping, and adopt a mutual exclusion strategy to guarantee a 4 KB separation b etween flips. We evaluate Flip-S on CIFAR-10 and ImageNet datasets across five V

iT architectures and two quantization levels. Results show that Flip-S achieves attack success rate (ASR) exceeding 90.0% on all models with 50 bits flipped, outperforming baselines with ASR typically below 80.0%. Furthermore, compared to the SOTA, Flip-S reduces the number of required bit-flips by 8x-20xwhile reaching equal or higher ASR. Our source code is publicly available.

****

Marten: Visual Question Answering with Mask Generation for Multi-modal Document Understanding

Zining Wang, Tongkun Guan, Pei Fu, Chen Duan, Qianyi Jiang, Zhentao Guo, Shan Guo, Junfeng Luo, Wei Shen, Xiaokang Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14460-14471

Multi-modal Large Language Models (MLLMs) have introduced a novel dimension to document understanding, i.e., they endow large language models with visual comprehension capabilities; however, how to design a suitable image-text pre-training task for bridging the visual and language modality in document-level MLLMs remains underexplored. In this study, we introduce a novel visual-language alignment method that casts the key issue as a Visual Question Answering with Mask generation (VQAMask) task, optimizing two tasks simultaneously: VQA-based text parsing and mask generation. The former allows the model to implicitly align images and text at the semantic level. The latter introduces an additional mask generator (discarded during inference) to explicitly ensure alignment between visual texts within images and their corresponding image regions at a spatially-aware level. Together, they can prevent model hallucinations when parsing visual text and effectively promote spatially-aware feature representation learning. To support the proposed VQAMask task, we construct a comprehensive image-mask generation pipeline and provide a large-scale dataset with 6M data (MTMask6M). Subsequently, we demonstrate that introducing the proposed mask generation task yields competitive document-level understanding performance. Leveraging the proposed VQAMask, we introduce Marten, a training-efficient MLLM tailored for document-level understanding. Extensive experiments show that our Marten consistently achieves significant improvements among 8B-MLLMs in document-centric tasks. Code and datasets are available at https://github.com/PriNing/Marten.

****

Mamba-Reg: Vision Mamba Also Needs Registers

Feng Wang, Jiahao Wang, Sucheng Ren, Guoyizhe Wei, Jieru Mei, Wei Shao, Yuyin Zhou, Alan Yuille, Cihang Xie; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14944-14953

Similar to Vision Transformers, this paper identifies artifacts also present within the feature maps of Vision Mamba. These artifacts, corresponding to high-norm tokens emerging in low-information background areas of images, appear much more severe in Vision Mamba---they exist prevalently even with the tiny-sized model and activate extensively across background regions. To mitigate this issue, we follow the prior solution of introducing register tokens into Vision Mamba. To better cope with Mamba blocks' uni-directional inference paradigm, two key modifications are introduced: 1) evenly inserting registers throughout the input token sequence, and 2) recycling registers for final decision predictions. We term this new architecture MambaReg. Qualitative observations suggest, compared to vanilla Vision Mamba, MambaReg's feature maps appear cleaner and more focused on semantically meaningful regions. Quantitatively, MambaReg attains stronger performance and scales better. For example, on the ImageNet benchmark, our MambaReg-B attains 83.0% accuracy, significantly outperforming Vim-B's 81.8%; furthermore, we provide the first successful scaling to the large model size (i.e., with 340M parameters), attaining a competitive accuracy of 83.6% (84.5% if finetuned with 384x384 inputs). Additional validation on the downstream semantic segmentation task also supports MambaReg's efficacy.

****

It's a (Blind) Match! Towards Vision-Language Correspondence without Parallel Data

Dominik Schnaus, Nikita Araslanov, Daniel Cremers; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24983-24992

The platonic representation hypothesis suggests that vision and language embeddings become more homogeneous as model and dataset sizes increase. In particular, pairwise distances within each modality become more similar. This suggests that as foundation models mature, it may become possible to match vision and language embeddings in a fully unsupervised fashion, i.e. without parallel data. We present the first feasibility study, and investigate conformity of existing vision and language foundation models in the context of unsupervised, or "blind", matching. First, we formulate unsupervised matching as a quadratic assignment problem and introduce a novel heuristic that outperforms previous solvers. We also develop a technique to find optimal matching problems, for which a non-trivial match is very likely. Second, we conduct an extensive study deploying a range of vision and language models on four datasets. Our analysis reveals that for many problem instances, vision and language representations can be indeed matched without supervision. This finding opens up the exciting possibility of embedding semantic knowledge into other modalities virtually annotation-free. As a proof of concept, we showcase an unsupervised classifier, which achieves non-trivial classification accuracy without any image-text annotation.

****

Open Set Label Shift with Test Time Out-of-Distribution Reference

Changkun Ye, Russell Tsuchida, Lars Petersson, Nick Barnes; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30619-30629

Open set label shift (OSLS) occurs when label distributions change from a source to a target distribution, and the target distribution has an additional out-of-distribution (OOD) class.In this work, we build estimators for both source and target open set label distributions using a source domain in-distribution (ID) classifier and an ID/OOD classifier. With reasonable assumptions on the ID/OOD classifier, the estimators are assembled into a sequence of three stages: 1) an estimate of the source label distribution of the OOD class, 2) an EM algorithm for Maximum Likelihood estimates (MLE) of the target label distribution, and 3) an estimate of the target label distribution of OOD class under relaxed assumptions on the OOD classifier.The sampling errors of estimates in 1) and 3) are quantified with a concentration inequality.The estimation result allows us to correct the ID classifier trained on the source distribution to the target distribution without retraining.Experiments on a variety of open set label shift settings demonstrate the effectiveness of our model.Our code is available at https://github.com/ChangkunYe/OpenSetLabelShift.

****

Visual Persona: Foundation Model for Full-Body Human Customization

Jisu Nam, Soowon Son, Zhan Xu, Jing Shi, Difan Liu, Feng Liu, Seungryong Kim, Yang Zhou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18630-18641

We introduce Visual Persona, a foundation model for text-to-image full-body human customization that, given a single in-the-wild human image, generates diverse images of the individual guided by text descriptions. Unlike prior methods that focus solely on preserving facial identity, our approach captures detailed full-body appearance, aligning with text descriptions for body structure and scene variations. Training this model requires large-scale paired human data, consisting of multiple images per individual with consistent full-body identities, which is notoriously difficult to obtain. To address this, we propose a data curation pipeline leveraging vision-language models to evaluate full-body appearance consistency, resulting in Visual Persona-500K--a dataset of 580k paired human images across 100k unique identities. For precise appearance transfer, we introduce a transformer encoder-decoder architecture adapted to a pre-trained text-to-image diffusion model, which augments the input image into distinct body regions, encodes these regions as local appearance features, and projects them into dense identity embeddings independently to condition the diffusion model for synthesizing customized images. Visual Persona consistently surpasses existing approaches, generating high-quality, customized images from in-the-wild inputs. Extensive ablation studies validate design choices, and we demonstrate the versatility of Visual Persona across various downstream tasks.

****
SOGS: Second-Order Anchor for Advanced 3D Gaussian Splatting
Jiahui Zhang, Fangneng Zhan, Ling Shao, Shijian Lu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11167-11176
Anchor-based 3D Gaussian splatting (3D-GS) exploits anchor features in 3D Gaussian prediction, which has achieved impressive 3D rendering quality with reduced Gaussian redundancy. On the other hand, it often encounters the dilemma among anchor features, model size, and rendering quality - large anchor features lead to large 3D models and high-quality rendering whereas reducing anchor features degrades Gaussian attribute prediction which leads to clear artifacts in the rendered textures and geometries. We design SOGS, an anchor-based 3D-GS technique that introduces second-order anchors to achieve superior rendering quality and reduced anchor features and model size simultaneously. Specifically, SOGS incorporates covariance-based second-order statistics and correlation across feature dimensions to augment features within each anchor, compensating for the reduced feature size and improving rendering quality effectively. In addition, it introduces a selective gradient loss to enhance the optimization of scene textures and scene geometries, leading to high-quality rendering with small anchor features. Extensive experiments over multiple widely adopted benchmarks show that SOGS achieves superior rendering quality in novel view synthesis with clearly reduced model size.
****
GaussianFormer-2: Probabilistic Gaussian Superposition for Efficient 3D Occupancy Prediction
Yuanhui Huang, Amonnut Thammatadatrakoon, Wenzhao Zheng, Yunpeng Zhang, Dalong Du, Jiwen Lu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27477-27486
3D semantic occupancy prediction has garnered attention as an important task for the robustness of vision-centric autonomous driving, which predicts fine-grained geometry and semantics of the surrounding scene. Most existing methods leverage dense grid-based scene representations, overlooking the spatial sparsity of the driving scenes, which leads to computational redundancy. Although 3D semantic Gaussian serves as an object-centric sparse alternative, most of the Gaussians still describe the empty region with low efficiency. To address this, we propose a probabilistic Gaussian superposition model which interprets each Gaussian as a probability distribution of its neighborhood being occupied and conforms to probabilistic multiplication to derive the overall geometry. Furthermore, we adopt the exact Gaussian mixture model for semantics calculation to avoid unnecessary overlapping of Gaussians. To effectively initialize Gaussians in non-empty region, we design a distribution-based initialization module which learns the pixel-aligned occupancy distribution instead of the depth of surfaces. We conduct extensive experiments on nuScenes and KITTI-360 datasets and our GaussianFormer-2 achieves state-of-the-art performance with high efficiency.
****
MExD: An Expert-Infused Diffusion Model for Whole-Slide Image Classification
Jianwei Zhao, Xin Li, Fan Yang, Qiang Zhai, Ao Luo, Yang Zhao, Hong Cheng, Huazhu Fu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20789-20799
Whole Slide Image (WSI) classification poses unique challenges due to the vast image size and numerous non-informative regions, which introduce noise and cause data imbalance during feature aggregation. To address these issues, we propose MExD, an Expert-Infused Diffusion Model that combines the strengths of a Mixture-of-Experts (MoE) mechanism with a diffusion model for enhanced classification. MExD balances patch feature distribution through a novel MoE-based aggregator that selectively emphasizes relevant information, effectively filtering noise, addressing data imbalance, and extracting essential features. These features are then integrated via a diffusion-based generative process to directly yield the class distribution for the WSI. Moving beyond conventional discriminative approaches, MExD represents the first generative strategy in WSI classification, capturing fine-grained details for robust and precise results. Our MExD is validated on t

hree widely-used benchmarks--Camelyon16, TCGA-NSCLC, and BRACS--consistently ach
ieving state-of-the-art performance in both binary and multi-class tasks.
****

Flexible Frame Selection for Efficient Video Reasoning
Shyamal Buch, Arsha Nagrani, Anurag Arnab, Cordelia Schmid; Proceedings of the C
omputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29071-29082
Video-language models have shown promise for addressing a range of multimodal ta
sks for video understanding, such as video question-answering. However, the inhe
rent computational challenges of processing long video data and increasing model
 sizes have led to standard approaches that are limited by the number of frames
they can process. In this work, we propose the Flexible Frame Selector (FFS), a
learnable policy model with a new flexible selection operation, that helps allev
iate input context restrictions by enabling video-language models to focus on th
e most informative frames for the downstream multimodal task, without adding und
ue processing cost. Our method differentiates from prior work due to its learnab
ility, efficiency, and flexibility. We verify the efficacy of our method on stan
dard video-question answering and reasoning benchmarks, and observe that our mod
el can improve base video-language model accuracy while reducing the number of d
ownstream processed frames.
****

EventGPT: Event Stream Understanding with Multimodal Large Language Models
Shaoyu Liu, Jianing Li, Guanghui Zhao, Yunjian Zhang, Xin Meng, Fei Richard Yu,
Xiangyang Ji, Ming Li; Proceedings of the Computer Vision and Pattern Recognitio
n Conference (CVPR), 2025, pp. 29139-29149
Event cameras capture visual information as asynchronous pixel change streams, e
xcelling in challenging lighting and high-dynamic scenarios. Existing multimodal
 large language models (MLLMs) concentrate on natural RGB images, failing in sce
narios where event data fits better. In this paper, we introduce EventGPT, the f
irst MLLM for event stream understanding, pioneering the integration of large la
nguage models (LLMs) with event-based vision. To bridge the huge domain gap, we
propose a three-stage optimization paradigm to progressively equip a pre-trained
 LLM with event understanding. Our EventGPT consists of an event encoder, a spat
io-temporal aggregator, a linear projector, an event-language adapter, and an LL
M. Firstly, GPT-generated RGB image-text pairs warm up the linear projector, fol
lowing LLaVA, as the gap between natural images and language is smaller. Secondl
y, we construct N-ImageNet-Chat, a large synthetic dataset of event data and cor
responding texts to enable the use of the spatio-temporal aggregator and to trai
n the event-language adapter, thereby aligning event features more closely with
the language space. Finally, we gather an instruction dataset, Event-Chat, which
 contains extensive real-world data to fine-tune the entire model, further enhan
cing its generalization ability. We construct a comprehensive benchmark, and exp
eriments show that EventGPT surpasses previous state-of-the-art MLLMs in generat
ion quality, descriptive accuracy, and reasoning capability.
****

Pixel-level and Semantic-level Adjustable Super-resolution: A Dual-LoRA Approach
Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, Lei Zhang; Pro
ceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025,
 pp. 2333-2343
Diffusion prior-based methods have shown impressive results in real-world image
super-resolution (SR). However, most existing methods entangle pixel-level and s
emantic-level SR objectives in the training process, struggling to balance pixel
-wise fidelity and perceptual quality. Meanwhile, users have varying preferences
 on SR results, thus it is demanded to develop an adjustable SR model that can b
e tailored to different fidelity-perception preferences during inference without
 re-training. We present Pixel-level and Semantic-level Adjustable SR (PiSA-SR),
 which learns two LoRA modules upon the pre-trained stable-diffusion (SD) model
to achieve improved and adjustable SR results. We first formulate the SD-based S
R problem as learning the residual between the low-quality input and the high-qu
ality output, then show that the learning objective can be decoupled into two di
stinct LoRA weight spaces: one is characterized by the l2-loss for pixel-level r

egression, and another is characterized by the LPIPS and classifier score distil lation losses to extract semantic information from pre-trained classification an d SD models. In its default setting, PiSA-SR can be performed in a single diffu sion step, achieving leading real-world SR results in both quality and efficienc y. By introducing two adjustable guidance scales on the two LoRA modules to cont rol the strengths of pixel-wise fidelity and semantic-level details during infer ence, PiSA-SR can offer flexible SR results according to user preference without re-training. The source code of our method can be found at https://github.com/c sslc/PiSA-SR.
****

Let Samples Speak: Mitigating Spurious Correlation by Exploiting the Clusterness of Samples
Weiwei Li, Junzhuo Liu, Yuanyuan Ren, Yuchen Zheng, Yahao Liu, Wen Li; Proceedin gs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1 5486-15496
Deep learning models are known to often learn features that spuriously correlate with the class label during training but are irrelevant to the prediction task. Existing methods typically address this issue by annotating potential spurious attributes, or filtering spurious features based on some empirical assumptions (e.g., simplicity of bias). However, these methods may yield unsatisfying perform ance due to the intricate and elusive nature of spurious correlations in real-wo rld data. In this paper, we propose a data-oriented approach to mitigate the spu rious correlation in deep learning models. We observe that samples that are infl uenced by spurious features tend to exhibit a dispersed distribution in the lear ned feature space. This allows us to identify the presence of spurious features. Subsequently, we obtain a bias-invariant representation by neutralizing the spu riousfeatures based on a simple grouping strategy. Then, we learn a feature tran sformation to eliminate the spuriousfeatures by aligning with this bias-invarian t representation. Finally, we update the classifier by incorporating the learned feature transformation and obtain an unbiased model. By integrating the aforeme ntioned identifying, neutralizing, eliminating and updating procedures, we build an effective pipeline for mitigating spurious correlation. Experiments on image and NLP debiasing benchmarks show an improvement in worst group accuracy of mor e than 20% compared to standard empirical risk minimization (ERM). Codes and ch eckpoints are available at https://github.com/davelee-uestc/nsf_debiasing.
****

Mr. DETR: Instructive Multi-Route Training for Detection Transformers
Chang-Bin Zhang, Yujie Zhong, Kai Han; Proceedings of the Computer Vision and Pa ttern Recognition Conference (CVPR), 2025, pp. 9933-9943
Existing methods enhance the training of detection transformers by incorporating an auxiliary one-to-many assignment. In this work, we treat the model as a mult i-task framework, simultaneously performing one-to-one and one-to-many predictio ns. We investigate the roles of each component in the transformer decoder across these two training targets, including self-attention, cross-attention, and feed -forward network. Our empirical results demonstrate that any independent compone nt in the decoder can effectively learn both targets simultaneously, even when o ther components are shared. This finding leads us to propose a multi-route train ing mechanism, featuring a primary route for one-to-one prediction and two auxil iary training routes for one-to-many prediction. We enhance the training mechani sm with a novel instructive self-attention that dynamically and flexibly guides object queries for one-to-many prediction. The auxiliary routes are removed duri ng inference, ensuring no impact on model architecture or inference cost. We con duct extensive experiments on various baselines, achieving consistent improvemen ts as shown in Fig. 1. Project page: https://visual-ai.github.io/mrdetr
****

MITracker: Multi-View Integration for Visual Object Tracking
Mengjie Xu, Yitao Zhu, Haotian Jiang, Jiaming Li, Zhenrong Shen, Sheng Wang, Hao lin Huang, Xinyu Wang, Han Zhang, Qing Yang, Qian Wang; Proceedings of the Compu ter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27176-27185
Multi-view object tracking (MVOT) offers promising solutions to challenges such

as occlusion and target loss, which are common in traditional single-view tracking. However, progress has been limited by the lack of comprehensive multi-view datasets and effective cross-view integration methods. To overcome these limitations, we compiled a Multi-View object Tracking (MVTrack) dataset of 234K high-quality annotated frames featuring 27 distinct objects across various scenes. In conjunction with this dataset, we introduce a novel MVOT method, Multi-View Integration Tracker (MITracker), to efficiently integrate multi-view object features and provide stable tracking outcomes. MITracker can track any object in video frames of arbitrary length from arbitrary viewpoints. The key advancements of our method over traditional single-view approaches come from two aspects: (1) MITracker transforms 2D image features into a 3D feature volume and compresses it into a bird's eye view (BEV) plane, facilitating inter-view information fusion; (2) we propose an attention mechanism that leverages geometric information from fused 3D feature volume to refine the tracking results at each view. MITracker outperforms existing methods on the MVTrack and GMTD datasets, achieving state-of-the-art performance.

****

Hearing Hands: Generating Sounds from Physical Interactions in 3D Scenes

Yiming Dou, Wonseok Oh, Yuqing Luo, Antonio Loquercio, Andrew Owens; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1795-1804

We study the problem of making 3D scene reconstructions interactive by asking the following question: can we predict the sounds of human hands physically interacting with a scene? First, we record a video of a human manipulating objects within a 3D scene using their hands. We then use these action-sound pairs to train a rectified flow model to map 3D hand trajectories to their corresponding audio. At test time, a user can query the model for other actions, parameterized as sequences of hand poses, to estimate their corresponding sounds. In our experiments, we find that our generated sounds accurately convey material properties and actions, and that they are often indistinguishable to human observers from real sounds. Project page: https://www.yimingdou.com/hearing_hands/.

****

AirRoom: Objects Matter in Room Reidentification

Runmao Yao, Yi Du, Zhuoqun Chen, Haoze Zheng, Chen Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1385-1394

Room reidentification (ReID) is a challenging yet essential task with numerous applications in fields such as augmented reality (AR) and homecare robotics. Existing visual place recognition (VPR) methods, which typically rely on global descriptors or aggregate local features, often struggle in cluttered indoor environments densely populated with man-made objects. These methods tend to overlook the crucial role of object-oriented information. To address this, we propose AirRoom, an object-aware pipeline that integrates multi-level object-oriented information--from global context to object patches, object segmentation, and keypoints--utilizing a coarse-to-fine retrieval approach. Extensive experiments on four newly constructed datasets--MPReID, HMReID, GibsonReID, and ReplicaReID--demonstrate that AirRoom outperforms state-of-the-art (SOTA) models across nearly all evaluation metrics, with improvements ranging from 6% to 80%. Moreover, AirRoom exhibits significant flexibility, allowing various modules within the pipeline to be substituted with different alternatives without compromising overall performance. It also shows robust and consistent performance under diverse viewpoint variations.

****

Not Only Text: Exploring Compositionality of Visual Representations in Vision-Language Models

Davide Berasi, Matteo Farina, Massimiliano Mancini, Elisa Ricci, Nicola Strisciuglio; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24917-24927

Vision-Language Models (VLMs) learn a shared feature space for text and images, enabling the comparison of inputs of different modalities. While prior works demonstrated that VLMs organize natural language representations into regular struc

tures encoding composite meanings, it remains unclear if compositional patterns also emerge in the visual embedding space. In this work, we investigate compositionality in the image domain, where the analysis of compositional properties is challenged by noise and sparsity of visual data. We address these problems and propose a framework, called Geodesically Decomposable Embeddings (GDE), that approximates image representations with geometry-aware compositional structures in the latent space. We demonstrate that visual embeddings of pre-trained VLMs exhibit a compositional arrangement, and evaluate the effectiveness of this property in the tasks of compositional classification and group robustness. GDE achieves stronger performance in compositional classification compared to its counterpart method that assumes linear geometry of the latent space. Notably, it is particularly effective for group robustness, where we achieve higher results than task-specific solutions. Our results indicate that VLMs can automatically develop a human-like form of compositional reasoning in the visual domain, making their underlying processes more interpretable. Code is available at https://github.com/BerasiDavide/vlm_image_compositionality.

****

DefMamba: Deformable Visual State Space Model

Leiye Liu, Miao Zhang, Jihao Yin, Tingwei Liu, Wei Ji, Yongri Piao, Huchuan Lu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8838-8847

Recently, state space models (SSM), particularly Mamba, have attracted significant attention from scholars due to their ability to effectively balance computational efficiency and performance. However, most existing visual Mamba methods flatten images into 1D sequences using predefined scan orders, which results the model being less capable of utilizing the spatial structural information of the image during the feature extraction process. To address this issue, we proposed a novel visual foundation model called DefMamba. This model includes a multi-scale backbone structure and deformable mamba (DM) blocks, which dynamically adjust the scanning path to prioritize important information, thus enhancing the capture and processing of relevant input features. By combining a deformable scanning (DS) strategy, this model significantly improves its ability to learn image structures and detects changes in object details. Numerous experiments have shown that DefMamba achieves state-of-the-art performance in various visual tasks, including image classification, object detection, instance segmentation, and semantic segmentation. The code is open source on DefMamba.

****

HOP: Heterogeneous Topology-based Multimodal Entanglement for Co-Speech Gesture Generation

Hongye Cheng, Tianyu Wang, Guangsi Shi, Zexing Zhao, Yanwei Fu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 906-916

Co-speech gestures are crucial non-verbal cues that enhance speech clarity and expressiveness in human communication, which have attracted increasing attention in multimodal research. While the existing methods have made strides in gesture accuracy, challenges remain in generating diverse and coherent gestures, as most approaches assume independence among multimodal inputs and lack explicit modeling of their interactions. In this work, we propose a novel multimodal learning method named HOP for co-speech gesture generation that captures the heterogeneous entanglement between gesture motion, audio rhythm, and text semantics, enabling the generation of coordinated gestures. By leveraging spatiotemporal graph modeling, we achieve the alignment of audio and action. Moreover, to enhance modality coherence, we build the audio-text semantic representation based on a reprogramming module, which is beneficial for cross-modality adaptation. Our approach enables the trimodal system to learn each other's features and represent them in the form of topological entanglement. Extensive experiments demonstrate that HOP achieves state-of-the-art performance, offering more natural and expressive co-speech gesture generation. More information, codes, and demos are available here: https://star-uu-wang.github.io/HOP/.

****

VoxelSplat: Dynamic Gaussian Splatting as an Effective Loss for Occupancy and Fl

ow Prediction
Ziyue Zhu, Shenlong Wang, Jin Xie, Jiang-jiang Liu, Jingdong Wang, Jian Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6761-6771
Recent advancements in camera-based occupancy prediction have focused on the simultaneous prediction of 3D semantics and scene flow, a task that presents significant challenges due to specific difficulties, e.g., occlusions and unbalanced dynamic environments. In this paper, we analyze these challenges and their underlying causes. To address them, we propose a novel regularization framework called VoxelSplat. This framework leverages recent developments in 3D Gaussian Splatting to enhance model performance in two key ways: (i)Enhanced Semantics Supervision through 2D Projection: During training, our method decodes sparse semantic 3D Gaussians from 3D representations and projects them onto the 2D camera view. This provides additional supervision signals in the camera-visible space, allowing 2D labels to improve the learning of 3D semantics. (ii) Scene Flow Learning: Our framework uses the predicted scene flow to model the motion of Gaussians, and is thus able to learn the scene flow of moving objects in a self-supervised manner using the labels of adjacent frames. Our method can be seamlessly integrated into various existing occupancy models, enhancingperformance without increasing inference time. Extensive experiments on benchmark datasets demonstrate the effectiveness of VoxelSplat in improving the accuracy of both semantic occupancy and scene flow estimation. The project page and codes are available at https://zzy816.github.io/VoxelSplat-Demo/.
****

Exploring Scene Affinity for Semi-Supervised LiDAR Semantic Segmentation
Chuandong Liu, Xingxing Weng, Shuguo Jiang, Pengcheng Li, Lei Yu, Gui-Song Xia; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27380-27389
This paper explores scene affinity (AIScene), namely intra-scene consistency and inter-scene correlation, for semi-supervised LiDAR semantic segmentation in driving scenes. Adopting teacher-student training, AIScene employs a teacher network to generate pseudo-labeled scenes from unlabeled data, which then supervise the student network's learning. Unlike most methods that include all points in pseudo-labeled scenes for forward propagation but only pseudo-labeled points for backpropagation, AIScene removes points without pseudo-labels, ensuring consistency in both forward and backward propagation within the scene. This simple point erasure strategy effectively prevents unsupervised, semantically ambiguous points (excluded in backpropagation) from affecting the learning of pseudo-labeled points. Moreover, AIScene incorporates patch-based data augmentation, mixing multiple scenes at both scene and instance levels. Compared to existing augmentation techniques that typically perform scene-level mixing between two scenes, our method enhances the semantic diversity of labeled (or pseudo-labeled) scenes, thereby improving the semi-supervised performance of segmentation models. Experiments show that AIScene outperforms previous methods on two popular benchmarks across four settings, achieving notable improvements of 1.9% and 2.1% in the most challenging 1% labeled data. The code will be released at https://github.com/azhuantou/AIScene.
****

ControlFace: Harnessing Facial Parametric Control for Face Rigging
Wooseok Jang, Youngjun Hong, Geonho Cha, Seungryong Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5614-5624
Manipulation of facial images to meet specific controls such as pose, expression, and lighting, also referred to as face rigging is a complex task in computer vision. Existing methods are limited by their reliance on image datasets, which necessitates individual-specific fine-tuning and limits their ability to retain fine-grained identity and semantic details, reducing practical usability. To overcome these limitations, we introduce ControlFace, a novel face rigging method conditioned on 3DMM renderings that enables flexible, high-fidelity control. ControlFace employs a dual-branch U-Nets: one, referred to as FaceNet, captures identity and fine details, while the other focuses on generation. To enhance control

precision, control mixer module encodes the correlated features between the target-aligned control and reference-aligned control, and a novel guidance method, reference control guidance, steers the generation process for better control adherence. By training on a facial video dataset, we fully utilize FaceNet's rich representations while ensuring control adherence. Extensive experiments demonstrate ControlFace's superior performance in identity preservation, and control precision, highlighting its practicality. Code and pre-trained weights will be publicly available.
****
Minority-Focused Text-to-Image Generation via Prompt Optimization
Soobin Um, Jong Chul Ye; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20926-20936
We investigate the generation of minority samples using pretrained text-to-image (T2I) latent diffusion models. Minority instances, in the context of T2I generation, can be defined as ones living on low-density regions of text-conditional data distributions. They are valuable for various applications of modern T2I generators, such as data augmentation and creative AI. Unfortunately, existing pretrained T2I diffusion models primarily focus on high-density regions, largely due to the influence of guided samplers (like CFG) that are essential for high-quality generation. To address this, we present a novel framework to counter the high-density-focus of T2I diffusion models. Specifically, we first develop an online prompt optimization framework that encourages emergence of desired properties during inference while preserving semantic contents of user-provided prompts. We subsequently tailor this generic prompt optimizer into a specialized solver that promotes generation of minority features by incorporating a carefully-crafted likelihood objective. Extensive experiments conducted across various types of T2I models demonstrate that our approach significantly enhances the capability to produce high-quality minority instances compared to existing samplers. Code is available at https://github.com/soobin-um/MinorityPrompt.
****
Imputation-free and Alignment-free: Incomplete Multi-view Clustering Driven by Consensus Semantic Learning
Yuzhuo Dai, Jiaqi Jin, Zhibin Dong, Siwei Wang, Xinwang Liu, En Zhu, Xihong Yang, Xinbiao Gan, Yu Feng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5071-5081
In incomplete multi-view clustering (IMVC), missing data induce prototype shifts within views and semantic inconsistencies across views. A feasible solution is to explore cross-view consistency in paired complete observations, further imputing and aligning the similarity relationships inherently shared across views. Nevertheless, existing methods are constrained by two-tiered limitations: (1) Neither instance- nor cluster-level consistency learning construct a semantic space shared across views to learn consensus semantics. The former enforces cross-view instances alignment, and wrongly regards unpaired observations with semantic consistency as negative pairs; the latter focuses on cross-view cluster counterparts while coarsely handling fine-grained intra-cluster relationships within views. (2) Excessive reliance on consistency results in unreliable imputation and alignment without incorporating view-specific cluster information. Thus, we propose an IMVC framework, imputation- and alignment-free for consensus semantics learning (FreeCSL). To bridge semantic gaps across all observations, we learn consensus prototypes from available data to discover a shared space, where semantically similar observations are pulled closer for consensus semantics learning. To capture semantic relationships within specific views, we design a heuristic graph clustering based on modularity to recover cluster structure with intra-cluster compactness and inter-cluster separation for cluster semantics enhancement. Extensive experiments demonstrate, compared to state-of-the-art competitors, FreeCSL achieves more confident and robust assignments on IMVC task.
****
Sensitivity-Aware Efficient Fine-Tuning via Compact Dynamic-Rank Adaptation
Tianran Chen, Jiarui Chen, Baoquan Zhang, Zhehao Yu, Shidong Chen, Rui Ye, Xutao Li, Yunming Ye; Proceedings of the Computer Vision and Pattern Recognition Conf

erence (CVPR), 2025, pp. 9655-9664
Parameter-Efficient Fine-Tuning (PEFT) is a fundamental research problem in comp uter vision, which aims to tune a few of parameters for efficient storage and ad aptation of pre-trained vision models. Recently, sensitivity-aware parameter eff icient fine-tuning method (SPT) addresses this problem by identifying sensitive parameters and then leveraging its sparse characteristic to combine unstructured and structured tuning for PEFT. However, existing methods only focus on sparse characteristic of sensitive parameters but overlook its distribution characteris tic, which results in additional storage burden and limited performance improvem ent. In this paper, we find that the distribution of sensitive parameters is not chaotic, but concentrates in a small number of rows or columns in each paramete r matrix. Inspired by this fact, we propose a Compact Dynamic-Rank Adaptation-ba sed tuning method for Sensitivity-aware Parameter efficient fine-Tuning, called CDRA-SPT. Specifically, we first identify the sensitive parameters that require tuning for each downstream task. Then, we reorganize the sensitive parameters by following its row and column into a compact sub-parameter matrix. Finally, a dy namic-rank adaptation is designed and applied at sub-parameter matrix level for PEFT. Its advantage is that the dynamic-rank characteristic of sub-parameter mat rix can be fully exploited for PEFT. Extensive experiments show that our method achieves superior performance over previous state-of-the-art methods.
****

MANTA: A Large-Scale Multi-View and Visual-Text Anomaly Detection Dataset for Ti ny Objects
Lei Fan, Dongdong Fan, Zhiguang Hu, Yiwen Ding, Donglin Di, Kai Yi, Maurice Pagn ucco, Yang Song; Proceedings of the Computer Vision and Pattern Recognition Conf erence (CVPR), 2025, pp. 25518-25527
We present MANTA, a visual-text anomaly detection dataset for tiny objects. The visual component comprises over 137.3K images across 38 object categories spanni ng five typical domains, of which 8.6K images are labeled as anomalous with pixe l-level annotations. Each image is captured from five distinct viewpoints to ens ure comprehensive object coverage. The text component consists of two subsets: D eclarative Knowledge, including 875 words that describe common anomalies across various domains and specific categories, with detailed explanations for \left< w hat, why, how \right>, including causes and visual characteristics; and Constru ctivist Learning, providing 2K multiple-choice questions with varying levels of difficulty, each paired with images and corresponded answer explanations. We als o propose a baseline for visual-text tasks and conduct extensive benchmarking ex periments to evaluate advanced methods across different settings, highlighting t he challenges and efficacy of our dataset.
****

MoDec-GS: Global-to-Local Motion Decomposition and Temporal Interval Adjustment for Compact Dynamic 3D Gaussian Splatting
Sangwoon Kwak, Joonsoo Kim, Jun Young Jeong, Won-Sik Cheong, Jihyong Oh, Munchur l Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CV PR), 2025, pp. 11338-11348
3D Gaussian Splatting (3DGS) has made significant strides in scene representatio n and neural rendering, with intense efforts focused on adapting it for dynamic scenes. Despite delivering remarkable rendering quality and speed, existing meth ods struggle with storage demands and the representation of complex real-world m otions. To address these challenges, we propose MoDec-GS, a memory-efficient Gau ssian splatting framework designed to reconstruct novel views in challenging sce narios with complex motions. We introduce Global-to-Local Motion Decomposition (GLMD) to effectively capture dynamic motions in a coarse-to-fine manner. This ap proach leverages Global Canonical Scaffolds (Global CS) and Local Canonical Scaf folds (Local CS), which extend static Scaffold representation to dynamic video r econstruction. For Global CS, we propose Global Anchor Deformation (GAD) to effi ciently represent global dynamics along complex motions by directly deforming th e implicit Scaffold attributes, including anchor position, offset, and local con text features. Next, we finely adjust local motions via the Local Gaussian Defor mation (LGD) of Local CS explicitly. Additionally, we introduce Temporal Interva

l Adjustment (TIA) to automatically control the temporal coverage of each Local CS during training, enabling MoDec-GS to find optimal interval assignments based on the specified number of temporal segments. Extensive evaluations demonstrate that MoDec-GS achieves an average 70% reduction in model size over state-of-the-art methods for dynamic 3D Gaussians from real-world dynamic videos while maintaining or even improving rendering quality.

****

DaCapo: Score Distillation as Stacked Bridge for Fast and High-quality 3D Editing

Yufei Huang, Bangyan Liao, Yuqi Hu, Haitao Lin, Lirong Wu, Siyuan Li, Cheng Tan, Zicheng Liu, Yunfan Liu, Zelin Zang, Chang Yu, Zhen Lei; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16304-16313

Score Distillation Sampling (SDS) has been successfully extended to text-driven 3D scene editing with 2D pretrained diffusion models. However, SDS-based editing methods suffer from lengthy optimization processes with slow inference and low quality. We attribute the issue of lengthy optimization to the stochastic optimization scheme used in SDS-based editing, where many steps may conflict with each other (e.g., the inherent trade-off between editing and preservation). To reduce this internal conflict and speed up the editing process, we propose to separate editing and preservation in time with a diffusion time schedule and frame the 3D editing optimization process as a diffusion bridge sampling process. Motivated by the analysis above, we introduce DaCapo, a fast diffusion sampling-like 3D editing method that incorporates a novel stacked bridge framework, which estimates a direct diffusion bridge between source and target distribution with only a pretrained 2D diffusion model. Specifically, It models the editing process as a combination of inversion and generation, where both processes happen simultaneously as a stack of Diffusion Bridges. DaCapo shows a 15x speed-up with comparable results to the state-of-the-art SDS-based method. It completes the process in just 2,500 steps on a single GPU and accommodates a variety of 3D representation methods.

****

A Selective Re-learning Mechanism for Hyperspectral Fusion Imaging

Yuanye Liu, Jinyang Liu, Renwei Dian, Shutao Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7437-7446

Hyperspectral fusion imaging is challenged by high computational cost due to the abundant spectral information. We find that pixels in regions with smooth spatial-spectral structure can be reconstructed well using a shallow network, while only those in regions with complex spatial-spectral structure require a deeper network. However, existing methods process all pixels uniformly, which ignores this property. To leverage this property, we propose a Selective Re-learning Fusion Network (SRLF) that initially extracts features from all pixels uniformly and then selectively refines distorted feature points. Specifically, SRLF first employs a Preliminary Fusion Module with robust global modeling capability to generate a preliminary fusion feature. Afterward, it applies a Selective Re-learning Module to focus on improving distorted feature points in the preliminary fusion feature. To achieve targeted learning, we present a novel Spatial-Spectral Structure-Guided Selective Re-learning Mechanism (SSG-SRL) that integrates the observation model to identify the feature points with spatial or spectral distortions. Only these distorted points are sent to the corresponding re-learning blocks, reducing both computational cost and the risk of overfitting. Finally, we develop an SRLF-Net, composed of multiple cascaded SRLFs, which surpasses multiple state-of-the-art methods on several datasets with minimal computational cost.

****

SCSegamba: Lightweight Structure-Aware Vision Mamba for Crack Segmentation in Structures

Hui Liu, Chen Jia, Fan Shi, Xu Cheng, Shengyong Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29406-29416

Pixel-level segmentation of structural cracks across various scenarios remains a considerable challenge. Current methods encounter challenges in effectively modeling crack morphology and texture, facing challenges in balancing segmentation

quality with low computational resource usage. To overcome these limitations, we propose a lightweight Structure-Aware Vision Mamba Network (SCSegamba), capable of generating high-quality pixel-level segmentation maps by leveraging both the morphological information and texture cues of crack pixels with minimal computational cost. Specifically, we developed a Structure-Aware Visual State Space module (SAVSS), which incorporates a lightweight Gated Bottleneck Convolution (GBC) and a Structure-Aware Scanning Strategy (SASS). The key insight of GBC lies in its effectiveness in modeling the morphological information of cracks, while the SASS enhances the perception of crack topology and texture by strengthening the continuity of semantic information between crack pixels. Experiments on crack benchmark datasets demonstrate that our method outperforms other state-of-the-art (SOTA) methods, achieving the highest performance with only 2.8M parameters. On the multi-scenario dataset, our method reached 0.8390 in F1 score and 0.8479 in mIoU.

****

Autoregressive Sequential Pretraining for Visual Tracking

Shiyi Liang, Yifan Bai, Yihong Gong, Xing Wei; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7254-7264

Recent advancements in visual object tracking have shifted towards a sequential generation paradigm, where object deformation and motion exhibit strong temporal dependencies. Despite the importance of these dependencies, widely adopted image-level pretrained backbones barely capture the dynamics in the consecutive video, which is the essence of tracking. Thus, we propose AutoRegressive Sequential Pretraining (ARP), an unsupervised spatio-temporal learner, via generating the evolution of object appearance and motion in video sequences.Our method leverages a diffusion model to autoregressively generate the future frame appearance, conditioned on historical embeddings extracted by a general encoder. Furthermore, to ensure trajectory coherence, the same encoder is employed to learn trajectory consistency by generating coordinate sequences in a reverse autoregressive fashion, a process we term back-tracking. Further, we integrate the pretrained ARP into ARTrackV2, creating ARPTrack, which is further fine-tuned for tracking tasks. ARPTrack achieves state-of-the-art performance across multiple benchmarks, becoming the first tracker to surpass 80% AO on GOT-10k, while maintaining high efficiency. These results demonstrate the effectiveness of our approach in capturing temporal dependencies for continuous video tracking.

****

Number it: Temporal Grounding Videos like Flipping Manga

Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, Xu Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13754-13765

Video Large Language Models (Vid-LLMs) have made remarkable advancements in comprehending video content for QA dialogue. However, they struggle to extend this visual understanding to tasks requiring precise temporal localization, known as Video Temporal Grounding (VTG). To address this, we introduce Number-Prompt (NumPro), a novel method that empowers Vid-LLMs to bridge visual comprehension with temporal grounding by adding unique numerical identifiers to each video frame. Treating a video as a sequence of numbered frame images, NumPro transforms VTG into an intuitive process: flipping through manga panels in sequence. This allows Vid-LLMs to "read" event timelines, accurately linking visual content with corresponding temporal information. Our experiments demonstrate that NumPro significantly boosts VTG performance of top-tier Vid-LLMs without additional computational cost. Furthermore, fine-tuning on a NumPro-enhanced dataset defines a new state-of-the-art for VTG, surpassing previous top-performing methods by up to 6.9% in mIoU for moment retrieval and 8.5% in mAP for highlight detection. The code is available at https://github.com/yongliang-wu/NumPro.

****

Collaborative Tree Search for Enhancing Embodied Multi-Agent Collaboration

Lizheng Zu, Lin Lin, Song Fu, Na Zhao, Pan Zhou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29513-29522

Embodied agents based on large language models (LLMs) face significant challenge

s in collaborative tasks, requiring effective communication and reasonable divis
ion of labor to ensure efficient and correct task completion. Previous approache
s with simple communication patterns carry erroneous or incoherent agent actions
, which can lead to additional risks. To address these problems, we propose Coop
erative Tree Search (CoTS), a framework designed to significantly improve collab
orative planning and task execution efficiency among embodied agents. CoTS guide
s multi-agents to discuss long-term strategic plans within a modified Monte Carl
o tree, searching along LLM-driven reward functions to provide a more thoughtful
 and promising approach to cooperation. Another key feature of our method is the
 introduction of a plan evaluation module, which not only prevents agent action
confusion caused by frequent plan updates but also ensures plan updates when the
 current plan becomes unsuitable. Experimental results show that the proposed me
thod performs excellently in planning, communication, and collaboration on embod
ied environments (CWAH and TDW-MAT), efficiently completing long-term, complex t
asks and significantly outperforming existing methods.
****

PromptHMR: Promptable Human Mesh Recovery
Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J. Black, Muhammed
 Kocabas; Proceedings of the Computer Vision and Pattern Recognition Conference
(CVPR), 2025, pp. 1148-1159
Human pose and shape (HPS) estimation presents challenges in diverse scenarios s
uch as crowded scenes, person-person interactions, and single-view reconstructio
n. Existing approaches lack mechanisms to incorporate auxiliary "side informatio
n" that could enhance reconstruction accuracy in such challenging scenarios. Fur
thermore, the most accurate methods rely on cropped person detections and cannot
 exploit scene context while methods that process the whole image often fail to
detect people and are less accurate than methods that use crops. While recent la
nguage-based methods explore HPS reasoning through large language or vision-lang
uage models, their metric accuracy is well below the state of the art. In contra
st, we present PromptHMR, a transformer-based promptable method that reformulate
s HPS estimation through spatial and semantic prompts. Our method processes full
 images to maintain scene context and accepts multiple input modalities: spatial
 prompts like bounding boxes and masks, and semantic prompts like language descr
iptions or interaction labels. PromptHMR demonstrates robust performance across
challenging scenarios: estimating people from bounding boxes as small as faces i
n crowded scenes, improving body shape estimation through language descriptions,
 modeling person-person interactions, and producing temporally coherent motions
in videos. Experiments on benchmarks show that PromptHMR achieves state-of-the-a
rt performance while offering flexible prompt-based control over the HPS estimat
ion process.
****

SyncVP: Joint Diffusion for Synchronous Multi-Modal Video Prediction
Enrico Pallotta, Sina Mokhtarzadeh Azar, Shuai Li, Olga Zatsarynna, Juergen Gall
; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR),
2025, pp. 13787-13797
Predicting future video frames is essential for decision-making systems, yet RGB
 frames alone often lack the information needed to fully capture the underlying
complexities of the real world. To address this limitation, we propose a multi-m
odal framework for Synchronous Video Prediction (SyncVP) that incorporates compl
ementary data modalities, enhancing the richness and accuracy of future predicti
ons. SyncVP builds on pre-trained modality-specific diffusion models and introdu
ces an efficient spatio-temporal cross-attention module to enable effective info
rmation sharing across modalities. We evaluate SyncVP on standard benchmark data
sets, such as Cityscapes and BAIR, using depth as an additional modality. We fur
thermore demonstrate its generalization to other modalities on SYNTHIA with sema
ntic information and ERA5-Land with climate data. Notably, SyncVP achieves state
-of-the-art performance, even in scenarios where only one modality is present, d
emonstrating its robustness and potential for a wide range of applications.
****

HUSH: Holistic Panoramic 3D Scene Understanding using Spherical Harmonics

Jongsung Lee, Harin Park, Byeong-Uk Lee, Kyungdon Joo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16599-16608
Motivated by the efficiency of spherical harmonics (SH) in representing various physical phenomena, we propose a Holistic panoramic 3D scene Understanding framework using Spherical Harmonics, dubbed as HUSH. Our approach focuses on a unified framework adaptable to various 3D scene understanding tasks via SH bases. To achieve this, we first estimate SH coefficients, allowing for the adaptive configuration of the SH bases specific to each scene. HUSH then employs a hierarchical attention module that uses SH bases as queries to generate comprehensive scene features by integrating these scene-adaptive SH bases with image features. Additionally, we introduce an SH basis index module that adaptively emphasizes relevant SH bases to produce task-relevant features, enhancing the versatility of HUSH across different scene understanding tasks. Finally, by combining the scene features with task-relevant features in the task-specific heads, we perform various scene understanding tasks, including depth, surface normal and room layout estimation. Experiments demonstrate that HUSH achieves state-of-the-art performance on depth estimation benchmarks, highlighting the robustness and scalability of using SH in panoramic 3D scene understanding.
****

SkillMimic: Learning Basketball Interaction Skills from Demonstrations
Yinhuai Wang, Qihan Zhao, Runyi Yu, Hok Wai Tsui, Ailing Zeng, Jing Lin, Zhengyi Luo, Jiwen Yu, Xiu Li, Qifeng Chen, Jian Zhang, Lei Zhang, Ping Tan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17540-17549
Traditional reinforcement learning methods for human-object interaction (HOI) rely on labor-intensive, manually designed skill rewards that do not generalize well across different interactions. We introduce SkillMimic, a unified data-driven framework that fundamentally changes how agents learn interaction skills by eliminating the need for skill-specific rewards. Our key insight is that a unified HOI imitation reward can effectively capture the essence of diverse interaction patterns from HOI datasets. This enables SkillMimic to learn a single policy that not only masters multiple interaction skills but also facilitates skill transitions, with both diversity and generalization improving as the HOI dataset grows. For evaluation, we collect and introduce two basketball datasets containing approximately 35 minutes of diverse basketball skills. Extensive experiments show that SkillMimic successfully masters a wide range of basketball skills including stylistic variations in dribbling, layup, and shooting. Moreover, these learned skills can be effectively composed by a high-level controller to accomplish complex and long-horizon tasks such as consecutive scoring, opening new possibilities for scalable and generalizable interaction skill learning. Project page: https://ingrid789.github.io/SkillMimic/
****

VISTREAM: Improving Computation Efficiency of Visual Streaming Perception via Law-of-Charge-Conservation Inspired Spiking Neural Network
Kang You, Ziling Wei, Jing Yan, Boning Zhang, Qinghai Guo, Yaoyu Zhang, Zhezhi He; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8796-8805
Visual streaming perception (VSP) involves online intelligent processing of sequential frames captured by vision sensors, enabling real-time decision-making in applications such as autonomous driving, UAVs, and AR/VR. However, the computational efficiency of VSP on edge devices remains a challenge due to power constraints and the underutilization of temporal dependencies between frames. While spiking neural networks (SNNs) offer biologically inspired event-driven processing with potential energy benefits, their practical advantage over artificial neural networks (ANNs) for VSP tasks remains unproven.In this work, we introduce a novel framework, VISTREAM, which leverages the Law of Charge Conservation (LoCC) property in ST-BIF neurons and a differential encoding (DiffEncode) scheme to optimize SNN inference for VSP. By encoding temporal differences between neighboring frames and eliminating frequent membrane resets, VISTREAM achieves significant computational efficiency while maintaining accuracy equivalent to its ANN counter

part. We provide theoretical proofs of equivalence and validate VISTREAM across diverse VSP tasks, including object detection, tracking, and segmentation, demonstrating substantial energy savings without compromising performance.

****

STPro: Spatial and Temporal Progressive Learning for Weakly Supervised Spatio-Temporal Grounding

Aaryan Garg, Akash Kumar, Yogesh S Rawat; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3384-3394

In this work, we study Weakly Supervised Spatio-Temporal Video Grounding (WSTVG), a challenging task of localizing subjects spatio-temporally in videos using only textual queries and no bounding box supervision. Inspired by recent advances in vision-language foundation models, we investigate their utility for WSTVG, leveraging their zero-shot grounding capabilities. However, we find that a simple adaptation lacks essential spatio-temporal grounding abilities. To bridge this gap, we introduce Tubelet Referral Grounding (TRG), which connects textual queries to tubelets to enable spatio-temporal predictions. Despite its promise, TRG struggles with compositional action understanding and dense scene scenarios. To address these limitations, we propose STPro, a progressive learning framework with two key modules: Sub-Action Temporal Curriculum Learning (SA-TCL), which incrementally builds compositional action understanding, and Congestion-Guided Spatial Curriculum Learning (CG-SCL), which adapts the model to complex scenes by spatially increasing task difficulty. STPro achieves state-of-the-art results on three benchmark datasets, with improvements of 1.0% on VidSTG-Declarative and 3.0% on HCSTVG-v1.

****

RGBAvatar: Reduced Gaussian Blendshapes for Online Modeling of Head Avatars

Linzhou Li, Yumeng Li, Yanlin Weng, Youyi Zheng, Kun Zhou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10747-10757

We present Reduced Gaussian Blendshapes Avatar (RGBAvatar), a method for reconstructing photorealistic, animatable head avatars at speeds sufficient for on-the-fly reconstruction. Unlike prior approaches that utilize linear bases from 3D morphable models (3DMM) to model Gaussian blendshapes, our method maps tracked 3DMM parameters into reduced blendshape weights with an MLP, leading to a compact set of blendshape bases. The learned compact base composition effectively captures essential facial details for specific individuals, and does not rely on the fixed base composition weights of 3DMM, leading to enhanced reconstruction quality and higher efficiency. To further expedite the reconstruction process, we develop a novel color initialization estimation method and a batch-parallel Gaussian rasterization process, achieving state-of-the-art quality with training throughput of about 630 images per second. Moreover, we propose a local-global sampling strategy that enables direct on-the-fly reconstruction, immediately reconstructing the model as video streams in real time while achieving quality comparable to offline settings. Our source code is available at https://github.com/gapszju/RGBAvatar.

****

Rashomon Sets for Prototypical-Part Networks: Editing Interpretable Models in Real-Time

Jon Donnelly, Zhicheng Guo, Alina Jade Barnett, Hayden McTavish, Chaofan Chen, Cynthia Rudin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4528-4538

Interpretability is critical for machine learning models in high-stakes settings because it allows users to verify the model's reasoning. In computer vision, prototypical part models (ProtoPNets) have become the dominant model type to meet this need. Users can easily identify flaws in ProtoPNets, but fixing problems in a ProtoPNet requires slow, difficult retraining that is not guaranteed to resolve the issue. This problem is called the "interaction bottleneck." We solve the interaction bottleneck for ProtoPNets by simultaneously finding many equally good ProtoPNets (i.e., a draw from a "Rashomon set"). We show that our framework - called Proto-RSet - quickly produces many accurate, diverse ProtoPNets, allowing users to correct problems in real time while maintaining performance guarantees

with respect to the training set. We demonstrate the utility of this method in two settings: 1) removing synthetic bias introduced to a bird-identification model and 2) debugging a skin cancer identification model. This tool empowers non-machine-learning experts, such as clinicians or domain experts, to quickly refine and correct machine learning models without repeated retraining by machine learning experts.

****

EEE-Bench: A Comprehensive Multimodal Electrical And Electronics Engineering Benchmark

Ming Li, Jike Zhong, Tianle Chen, Yuxiang Lai, Konstantinos Psounis; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13337-13349

Recent studies on large language models (LLMs) and large multimodal models (LMMs) have demonstrated promising skills in various domains including science and mathematics. However, their capability in more challenging and real-world related scenarios like engineering has not been systematically studied. To bridge this gap, we propose EEE-Bench, a multimodal benchmark aimed at assessing LMMs' capabilities in solving practical engineering tasks, using electrical and electronics engineering (EEE) as the testbed. Our benchmark consists of 2860 hand-picked and carefully curated problems spanning 10 essential subdomains such as analog circuits, control systems, etc. Compared to other domains, engineering problems are intrinsically 1) more visually complex and versatile and 2) less deterministic in solutions. Successful solutions to these problems often demand more-than-usual rigorous integration of visual and textual information as models need to understand intricate images like abstract circuits and system diagrams while taking professional instructions. Alongside EEE-Bench, we provide extensive quantitative evaluations, fine-grained analysis, and improvement methods using 17 widely-used open- and closed-sourced LLMs and LMMs and 7 popular prompting techniques. Our results reveal notable deficiencies in current foundation models for EEE, including an average performance ranging from 19.48% to 46.78% and a tendency toward "laziness" in overlooking essential visual context. In summary, we believe EEE-Bench not only reveals some noteworthy limitations of LMMs but also provides a valuable resource for advancing research on their application in practical engineering tasks, driving future improvements in their capability to handle complex, real-world scenarios.

****

Text-Driven Fashion Image Editing with Compositional Concept Learning and Counterfactual Abduction

Shanshan Huang, Haoxuan Li, Chunyuan Zheng, Mingyuan Ge, Wei Gao, Lei Wang, Li Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28726-28735

Fashion image editing is a valuable tool for designers to convey their creative ideas by visualizing design concepts. With the recent advances in text editing methods, significant progress has been made in fashion image editing. However, they face two key challenges: spurious correlations in training data often induce changes in other areas when editing an area representing the intended editing concept, and these models typically lack the ability to edit multiple concepts simultaneously. To address the above challenges, we propose a novel \underline T ext-driven \underline F ashion \underline I mage edi\underline T ing framework called T-FIT to mitigate the impact of spurious correlation by integrating counterfactual reasoning with compositional concept learning to precisely ensure compositional multi-concept fashion image editing relying solely on text descriptions. Specifically, T-FIT includes three key components. (i) Counterfactual abduction module, which learns an exogenous variable of the source image by a denoising U-Net model. (ii) Concept learning module, which identifies concepts in fashion image editing--such as clothing types and colors and projects a target concept into the space spanned from a series of textual prompts. (iii) Concept composition module, which enables simultaneous adjustments of multiple concepts by aggregating each concept's direction vector obtained from the concept learning module. Extensive experiments show that our method can achieve state-of-the-art performanc

e on various fashion image editing tasks, including single-concept editing (e.g. , sleeve length, clothing type) and multi-concept editing (e.g., color & sleeve length).
****

EnvPoser: Environment-aware Realistic Human Motion Estimation from Sparse Observ ations with Uncertainty Modeling
Songpengcheng Xia, Yu Zhang, Zhuo Su, Xiaozheng Zheng, Zheng Lv, Guidong Wang, Y ongjie Zhang, Qi Wu, Lei Chu, Ling Pei; Proceedings of the Computer Vision and P attern Recognition Conference (CVPR), 2025, pp. 1839-1849
Estimating full-body motion using the tracking signals of head and hands from VR devices holds great potential for various applications. However, the sparsity a nd unique distribution of observations present a significant challenge, resultin g in an ill-posed problem with multiple feasible solutions (i.e., hypotheses). T his amplifies uncertainty and ambiguity in full-body motion estimation, especial ly for the lower-body joints. Therefore, we propose a new method, EnvPoser, that employs a two-stage framework to perform full-body motion estimation using spar se tracking signals and pre-scanned environment from VR devices. EnvPoser models the multi-hypothesis nature of human motion through an uncertainty-aware estima tion module in the first stage. In the second stage, we refine these multi-hypot hesis estimates by integrating semantic and geometric environmental constraints, ensuring that the final motion estimation aligns realistically with both the en vironmental context and physical interactions.Qualitative and quantitative exper iments on two public datasets demonstrate that our method achieves state-of-the- art performance, highlighting significant improvements in human motion estimatio n within motion-environment interaction scenarios. Project page: https://xspc.gi thub.io/EnvPoser/.
****

A Unified Framework for Heterogeneous Semi-supervised Learning
Marzi Heidari, Abdullah Alchihabi, Hao Yan, Yuhong Guo; Proceedings of the Compu ter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15371-15380
In this work, we introduce a novel problem setup termed as Heterogeneous Semi-Su pervised Learning (HSSL), which presents unique challenges by bridging the semi- supervised learning (SSL) task and the unsupervised domain adaptation (UDA) task , and expanding standard semi-supervised learning to cope with heterogeneous tra ining data. At its core, HSSL aims to learn a prediction model using a combinati on of labeled and unlabeled training data drawn separately from heterogeneous do mains that share a common set of semantic categories; this model is intended to differentiate the semantic categories of test instances sampled from both the la beled and unlabeled domains. In particular, the labeled and unlabeled domains ha ve dissimilar label distributions and class feature distributions. This heteroge neity, coupled with the assorted sources of the test data, introduces significan t challenges to standard SSL and UDA methods. Therefore, we propose a novel meth od, Unified Framework for Heterogeneous Semi-supervised Learning (Uni-HSSL), to address HSSL by directly learning a fine-grained classifier from the heterogeneo us data, which adaptively handles the inter-domain heterogeneity while leveragin g both the unlabeled data and the inter-domain semantic class relationships for cross-domain knowledge transfer and adaptation. We conduct comprehensive experim ents and the experimental results validate the efficacy and superior performance of the proposed Uni-HSSL over state-of-the-art semi-supervised learning and uns upervised domain adaptation methods.
****

Adapting Text-to-Image Generation with Feature Difference Instruction for Generi c Image Restoration
Chao Wang, Hehe Fan, Huichen Yang, Sarvnaz Karimi, Lina Yao, Yi Yang; Proceeding s of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23 539-23550
Diffusion-based Text-to-Image (T2I) models have demonstrated significant potenti al in image restoration. However, existing models continue to grapple with chall enges such as complex training and prompt design. We introduce a new perspective for improving image restoration by injecting knowledge from pretrained vision-l

anguage models into current T2I models. We empirically show that the degradation and content representations in BLIP-2 can be linearly separated, providing promising degradation guidance for image restoration. Specifically, the Feature Difference Instruction (FDI) is first extracted by Q-Formers through a simple subtraction operation based on reference image pairs. Then, we propose a multi-scale FDI adapter to decouple the degradation style and corrupted artifacts, and inject the styleflow exclusively into specific blocks through adapter-tuning, thereby preventing noise interference and eschewing the need for cumbersome weight retraining. In this way, we can train various task-specific adapters according to different degradations, achieving rich detail enhancement in the restoration results. Furthermore, the proposed FDI adapters have attractive properties of practical value, such as composability and generalization ability for all-in-one and mixed-degradation restoration. Extensive experiments under various settings demonstrate that our method has promising repairing quality over 10 image restoration tasks and a wide range of other applications.

****

ReCon: Enhancing True Correspondence Discrimination through Relation Consistency for Robust Noisy Correspondence Learning

Quanxing Zha, Xin Liu, Shu-Juan Peng, Yiu-ming Cheung, Xing Xu, Nannan Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29680-29689

Can we accurately identify the true correspondences from multimodal datasets containing mismatched data pairs? Existing methods primarily emphasize the similarity matching between the representations of objects across modalities, potentially neglecting the crucial relation consistency within modalities that are particularly important for distinguishing the true and false correspondences. Such an omission often runs the risk of misidentifying negatives as positives, thus leading to unanticipated performance degradation. To address this problem, we propose a general Relation Consistency learning framework, namely ReCon, to accurately discriminate the true correspondences among the multimodal data and thus effectively mitigate the adverse impact caused by mismatches. Specifically, ReCon leverages a novel relation consistency learning to ensure the dual-alignment, respectively of, the cross-modal relation consistency between different modalities and the intra-modal relation consistency within modalities. Thanks to such dual constrains on relations, ReCon significantly enhances its effectiveness for true correspondence discrimination and therefore reliably filters out the mismatched pairs to mitigate the risks of wrong supervisions. Extensive experiments on three widely-used benchmark datasets, including Flickr30K, MS-COCO, and Conceptual Captions, are conducted to demonstrate the effectiveness and superiority of ReCon compared with other SOTAs. The code is available at: https://anonymous.4open.science/r/ReCon-NCL.

****

Free360: Layered Gaussian Splatting for Unbounded 360-Degree View Synthesis from Extremely Sparse and Unposed Views

Chong Bao, Xiyu Zhang, Zehao Yu, Jiale Shi, Guofeng Zhang, Songyou Peng, Zhaopeng Cui; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16377-16387

Neural rendering has demonstrated remarkable success in high-quality 3D neural reconstruction and novel view synthesis with dense input views and accurate poses. However, applying it to sparse, unposed views in unbounded 360* scenes remains a challenging problem. In this paper, we propose a novel neural rendering framework to accomplish the unposed and extremely sparse-view 3D reconstruction in unbounded 360* scenes. To resolve the spatial ambiguity inherent in unbounded scenes with sparse input views, we propose a layered Gaussian-based representation to effectively model the scene with distinct spatial layers. By employing a dense stereo reconstruction model to recover coarse geometry, we introduce a layer-specific bootstrap optimization to refine the noise and fill occluded regions in the reconstruction. Furthermore, we propose an iterative fusion of reconstruction and generation alongside an uncertainty-aware training approach to facilitate mutual conditioning and enhancement between these two processes. Comprehensive ex

periments show that our approach outperforms existing state-of-the-art methods i
n terms of rendering quality and surface reconstruction accuracy. Project page:
https://zju3dv.github.io/free360/
****

Tuning the Frequencies: Robust Training for Sinusoidal Neural Networks
Tiago Novello, Diana Aldana, Andre Araujo, Luiz Velho; Proceedings of the Comput
er Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3071-3080
Sinusoidal neural networks have been shown effective as implicit neural represen
tations (INRs) of low-dimensional signals, due to their smoothness and high repr
esentation capacity. However, initializing and training them remain empirical ta
sks which lack on deeper understanding to guide the learning process. To fill th
is gap, our work introduces a theoretical framework that explains the capacity p
roperty of sinusoidal networks and offers robust control mechanisms for initiali
zation and training. Our analysis is based on a novel amplitude-phase expansion
of the sinusoidal multilayer perceptron, showing how its layer compositions prod
uce a large number of new frequencies expressed as integer combinations of the i
nput frequencies. This relationship can be directly used to initialize the input
neurons, as a form of spectral sampling, and to bound the network's spectrum wh
ile training. Our method, referred to as TUNER (TUNing sinusoidal nEtwoRks), gre
atly improves the stability and convergence of sinusoidal INR training, leading
to detailed reconstructions, while preventing overfitting.
****

Preconditioners for the Stochastic Training of Neural Fields
Shin-Fang Chng, Hemanth Saratchandran, Simon Lucey; Proceedings of the Computer
Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27222-27232
Neural fields encode continuous multidimensional signals as neural networks, ena
bling diverse applications in computer vision, robotics, and geometry. While Ada
m is effective for stochastic optimization, it often requires long training time
s. To address this, we explore alternative optimization techniques to accelerate
training without sacrificing accuracy. Traditional second-order methods like L-
BFGS are unsuitable for stochastic settings. We propose a theoretical framework
for training neural fields with curvature-aware diagonal preconditioners, demons
trating their effectiveness across tasks such as image reconstruction, shape mod
eling, and Neural Radiance Fields (NeRF).
****

Open Ad-hoc Categorization with Contextualized Feature Learning
Zilin Wang, Sangwoo Mo, Stella X. Yu, Sima Behpour, Liu Ren; Proceedings of the
Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15108-15117
Adaptive categorization of visual scenes is essential for AI agents to handle ch
anging tasks. Unlike fixed common categories for plants or animals, ad-hoc cate
gories, such as things to sell at a garage sale, are created dynamically to achi
eve specific tasks. We study open ad-hoc categorization, where the goal is to i
nfer novel concepts and categorize images based on a given context, a small set
of labeled exemplars, and some unlabeled data. We have two key insights: 1) reco
gnizing ad-hoc categories relies on the same perceptual processes as common cate
gories; 2) novel concepts can be discovered semantically by expanding contextual
cues or visually by clustering similar patterns. We propose OAK, a simple mode
l that introduces a single learnable context token into CLIP, trained with CLIP'
s objective of aligning visual and textual features and GCD's objective of clust
ering similar images. On Stanford and Clevr-4 datasets, OAK consistently achieve
s the state-of-art in accuracy and concept discovery across multiple categorizat
ions, including 87.4% novel accuracy on Stanford Mood, surpassing CLIP and GCD b
y over 50%. Moreover, OAK generates interpretable saliency maps, focusing on han
ds for Action, faces for Mood, and backgrounds for Location, promoting transpare
ncy and trust while enabling accurate and flexible categorization.
****

Real-time Free-view Human Rendering from Sparse-view RGB Videos using Double Unp
rojected Textures
Guoxing Sun, Rishabh Dabral, Heming Zhu, Pascal Fua, Christian Theobalt, Marc Ha
bermann; Proceedings of the Computer Vision and Pattern Recognition Conference (

CVPR), 2025, pp. 562-573
Real-time free-view human rendering from sparse-view RGB inputs is a challenging task due to the sensor scarcity and the tight time budget. To ensure efficiency, recent methods leverage 2D CNNs operating in texture space to learn rendering primitives. However, they either jointly learn geometry and appearance, or completely ignore sparse image information for geometry estimation, significantly harming visual quality and robustness to unseen body poses. To address these issues, we present Double Unprojected Textures, which at the core disentangles coarse geometric deformation estimation from appearance synthesis, enabling robust and photorealistic 4K rendering in real-time. Specifically, we first introduce a novel image-conditioned template deformation network, which estimates the coarse deformation of the human template from a first unprojected texture. This updated geometry is then used to apply a second and more accurate texture unprojection. The resulting texture map has fewer artifacts and better alignment with input views, which benefits our learning of finer-level geometry and appearance represented by Gaussian splats. We validate the effectiveness and efficiency of the proposed method in quantitative and qualitative experiments, which significantly surpasses other state-of-the-art methods.
****
ECBench: Can Multi-modal Foundation Models Understand the Egocentric World? A Holistic Embodied Cognition Benchmark
Ronghao Dang, Yuqian Yuan, Wenqi Zhang, Yifei Xin, Boqiang Zhang, Long Li, Liuyi Wang, Qinyang Zeng, Xin Li, Lidong Bing; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24593-24602
The enhancement of generalization in robots by large vision-language models (LVLMs) is increasingly evident. Therefore, the embodied cognitive abilities of LVLMs based on egocentric videos are of great interest. However, current datasets for embodied video question answering lack comprehensive and systematic evaluation frameworks. Critical embodied cognitive issues, such as robotic self-cognition, dynamic scene perception, and hallucination, are rarely addressed. To tackle these challenges, we propose ECBench, a high-quality benchmark designed to systematically evaluate the embodied cognitive abilities of LVLMs. ECBench features a diverse range of scene video sources, open and varied question formats, and 30 dimensions of embodied cognition. To ensure quality, balance, and high visual dependence, ECBench uses class-independent meticulous human annotation and multi-round question screening strategies. Additionally, we introduce ECEval, a comprehensive evaluation system that ensures the fairness and rationality of the indicators. Utilizing ECBench, we conduct extensive evaluations of proprietary, open-source, and task-specific LVLMs. ECBench is pivotal in advancing the embodied cognitive capabilities of LVLMs, laying a solid foundation for developing reliable core models for embodied agents. All data and code is available at https://github.com/Rh-Dang/ECBench.
****
SfM-Free 3D Gaussian Splatting via Hierarchical Training
Bo Ji, Angela Yao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21654-21663
Standard 3D Gaussian Splatting (3DGS) relies on known or pre-computed camera poses and a sparse point cloud, obtained from structure-from-motion (SfM) preprocessing, to initialize and grow 3D Gaussians. We propose a novel SfM-Free 3DGS (SFGS) method for video input, eliminating the need for known camera poses and SfM preprocessing. Our approach introduces a hierarchical training strategy that trains and merges multiple 3D Gaussian representations -- each optimized for specific scene regions -- into a single, unified 3DGS model representing the entire scene. To compensate for large camera motions, we leverage video frame interpolation models. Additionally, we incorporate multi-source supervision to reduce overfitting and enhance representation. Experimental results reveal that our approach significantly surpasses state-of-the-art SfM-free novel view synthesis methods. On the Tanks and Temples dataset, we improve PSNR by an average of 2.25dB, with a maximum gain of 3.72dB in the best scene. On the CO3D-V2 dataset, we achieve an average PSNR boost of 1.74dB, with a top gain of 3.90dB. The code is available

at \href https://github.com/jibo27/3DGS_Hierarchical_Training/ https://github.com/jibo27/3DGS_Hierarchical_Training .
****

Large Self-Supervised Models Bridge the Gap in Domain Adaptive Object Detection
Marc-Antoine Lavoie, Anas Mahmoud, Steven L. Waslander; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4692-4702
The current state-of-the-art methods in domain adaptive object detection (DAOD) use Mean Teacher self-labelling, where a teacher model, directly derived as an exponential moving average of the student model, is used to generate labels on the target domain which are then used to improve both models in a positive loop. This couples learning and generating labels on the target domain, and other recent works also leverage the generated labels to add additional domain alignment losses. We believe this coupling is brittle and excessively constrained: there is no guarantee that a student trained only on source data can generate accurate target domain labels and initiate the positive feedback loop, and much better target domain labels can likely be generated by using a large pretrained network that has been exposed to much more data. Vision foundational models are exactly such models, and they have shown impressive task generalization capabilities even when frozen. We want to leverage these models for DAOD and introduce DINO Teacher, which consists of two components. First, we train a new labeller on source data only using a large frozen DINOv2 backbone and show it generates more accurate labels than Mean Teacher. Next, we align the student's source and target image patch features with those from a DINO encoder, driving source and target representations closer to the generalizable DINO representation. We obtain state-of-the-art performance on multiple DAOD datasets.
****

Dynamic Updates for Language Adaptation in Visual-Language Tracking
Xiaohai Li, Bineng Zhong, Qihua Liang, Zhiyi Mo, Jian Nong, Shuxiang Song; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19165-19174
The consistency between the semantic information provided by the multi-modal reference and the tracked object is crucial for visual-language (VL) tracking. However, existing VL tracking frameworks rely on static multi-modal references to locate dynamic objects, which can lead to semantic discrepancies and reduce the robustness of the tracker. To address this issue, we propose a novel vision-language tracking framework, named DUTrack, which captures the latest state of the target by dynamically updating multi-modal references to maintain consistency.Specifically, we introduce a Dynamic Language Update Module, which leverages a large language model to generate dynamic language descriptions for the object based on visual features and object category information. Then, we design a Dynamic Template Capture Module, which captures the regions in the image that highly match the dynamic language descriptions. Furthermore, to ensure the efficiency of description generation, we design an update strategy that assesses changes in target displacement, scale, and other factors to decide on updates. Finally, the dynamic template and language descriptions that record the latest state of the target are used to update the multi-modal references, providing more accurate reference information for subsequent inference and enhancing the robustness of the tracker.DUTrack achieves new state-of-the-art performance on four mainstream vision-language and two vision-only tracking benchmarks, including LaSOT, LaSOT_ext, TNL2K, OTB99-Lang, GOT-10K, and UAV123.
****

Multi-focal Conditioned Latent Diffusion for Person Image Synthesis
Jiaqi Liu, Jichao Zhang, Paolo Rota, Nicu Sebe; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16019-16028
The Latent Diffusion Model (LDM) has demonstrated strong capabilities in high-resolution image generation and has been widely employed for Pose-Guided Person Image Synthesis (PGPIS), yielding promising results. However, the compression process of LDM often results in the deterioration of details, particularly in sensitive areas such as facial features and clothing textures. In this paper, we propose a Multi-focal Conditioned Latent Diffusion (MCLD) method to address these lim

itations by conditioning the model on disentangled, pose-invariant features from these sensitive regions. Our approach utilizes a multi-focal condition aggregation module, which effectively integrates facial identity and texture-specific information, enhancing the model's ability to produce appearance realistic and identity-consistent images. Our method demonstrates consistent identity and appearance generation on the DeepFashion dataset and enables flexible person image editing due to its generation consistency. The code is available at https://github.com/jqliu09/mcld.

****

Uncertainty Meets Diversity: A Comprehensive Active Learning Framework for Indoor 3D Object Detection

Jiangyi Wang, Na Zhao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20329-20339

Active learning has emerged as a promising approach to reduce the substantial annotation burden in 3D object detection tasks, spurring several initiatives in outdoor environments. However, its application in indoor environments remains unexplored. Compared to outdoor 3D datasets, indoor datasets face significant challenges, including fewer training samples per class, a greater number of classes, more severe class imbalance, and more diverse scene types and intra-class variances.This paper presents the first study on active learning for indoor 3D object detection, where we propose a novel framework tailored for this task. Our method incorporates two key criteria - uncertainty and diversity - to actively select the most ambiguous and informative unlabeled samples for annotation. The uncertainty criterion accounts for both inaccurate detections and undetected objects, ensuring that the most ambiguous samples are prioritized. Meanwhile, the diversity criterion is formulated as a joint optimization problem that maximizes the diversity of both object class distributions and scene types, using a new Class-aware Adaptive Prototype (CAP) bank. The CAP bank dynamically allocates representative prototypes to each class, helping to capture varying intra-class diversity across different categories.We evaluate our method on SUN RGB-D and ScanNetV2, where it outperforms baselines by a significant margin, achieving over 85% of fully-supervised performance with just 10% of the annotation budget.

****

CASAGPT: Cuboid Arrangement and Scene Assembly for Interior Design

Weitao Feng, Hang Zhou, Jing Liao, Li Cheng, Wenbo Zhou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29173-29182

We present a novel approach for indoor scene synthesis, which learns to arrange decomposed cuboid primitives to represent 3D objects within a scene. Unlike conventional methods that use bounding boxes to determine the placement and scale of 3D objects, our approach leverages cuboids as a straightforward yet highly effective alternative for modeling objects. This allows for compact scene generation while minimizing object intersections. Our approach, coined CASAGPT for Cuboid Arrangement and Scene Assembly, employs an autoregressive model to sequentially arrange cuboids, producing physically plausible scenes. By applying rejection sampling during the fine-tuning stage to filter out scenes with object collisions, our model further reduces intersections and enhances scene quality. Additionally, we introduce a refined dataset, 3DFRONT-NC, which eliminates significant noise presented in the original dataset, 3D-FRONT. Extensive experiments on the 3D-FRONT dataset as well as our dataset demonstrate that our approach consistently outperforms the state-of-the-art methods, enhancing the realism of generated scenes, and providing a promising direction for 3D scene synthesis.

****

Identity-Clothing Similarity Modeling for Unsupervised Clothing Change Person Re-Identification

Zhiqi Pang, Junjie Wang, Lingling Zhao, Chunyu Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19251-19260

Clothing change person re-identification (CC-ReID) aims to match different images of the same person, even when the clothing varies across images. To reduce manual labeling costs, existing unsupervised CC-ReID methods employ clustering algorithms to generate pseudo-labels. However, they often fail to assign the same ps

eudo-label to two images with the same identity but different clothing--referred to as a clothing change positive pair--thus hindering clothing-invariant feature learning. To address this issue, we propose the identity-clothing similarity modeling (ICSM) framework. To effectively connect clothing change positive pairs, ICSM first performs clothing-aware learning to leverage all discriminative information, including clothing, to obtain compact clusters. It then extracts cluster-level identity and clothing features and performs inter-cluster similarity estimation to identify clothing change positive clusters, reliable negative clusters, and hard negative clusters for each compact cluster. During optimization, we design an adaptive version of existing optimization methods to enhance similarities of clothing change positive pairs, while also introducing text semantics as a supervisory signal to further promote clothing invariance. Extensive experimental results across multiple datasets validate the effectiveness of the proposed framework, demonstrating its superiority over existing unsupervised methods and its competitiveness with some supervised approaches.

****

Evaluating Model Perception of Color Illusions in Photorealistic Scenes

Lingjun Mao, Zineng Tang, Alane Suhr; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7805-7814

We study the perception of color illusions by vision-language models. Color illusion, where a person's visual system perceives color differently from actual color, is well-studied in human vision. However, it remains underexplored whether vision-language models (VLMs), trained on large-scale human data, exhibit similar perceptual biases when confronted with such color illusions. We propose an automated framework for generating color illusion images, resulting in RCID (Realistic Color Illusion Dataset), a dataset of 19,000 realistic illusion images. Our experiments show that all studied VLMs exhibit perceptual biases similar human vision. Finally, we train a model to distinguish both human perception and actual pixel differences.

****

MINIMA: Modality Invariant Image Matching

Jiangwei Ren, Xingyu Jiang, Zizhuo Li, Dingkang Liang, Xin Zhou, Xiang Bai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23059-23068

Image matching for both cross-view and cross-modality plays a critical role in multimodal perception. In practice, the modality gap caused by different imaging systems/styles poses great challenges to the matching task. Existing works try to extract invariant features for specific modalities and train on limited datasets, showing poor generalization. In this paper, we present MINIMA, a unified image matching framework for multiple cross-modal cases. Without pursuing fancy modules, our MINIMA aims to enhance universal performance from the perspective of data scaling up. For such purpose, we propose a simple yet effective data engine that can freely produce a large dataset containing multiple modalities, rich scenarios, and accurate matching labels. Specifically, we scale up the modalities from cheap but rich RGB-only matching data, by means of generative models. Under this setting, the matching labels and rich diversity of the RGB dataset are well inherited by the generated multimodal data. Benefiting from this, we construct MD-syn, a new comprehensive dataset that fills the data gap for general multimodal image matching. With MD-syn, we can directly train any advanced matching pipeline on randomly selected modality pairs to obtain cross-modal ability. Extensive experiments on in-domain and zero-shot matching tasks, including 19 cross-modal cases, demonstrate that our MINIMA can significantly outperform the baselines and even surpass modality-specific methods. The dataset and code are available at https://github.com/LSXI7/MINIMA

****

OccMamba: Semantic Occupancy Prediction with State Space Models

Heng Li, Yuenan Hou, Xiaohan Xing, Yuexin Ma, Xiao Sun, Yanyong Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11949-11959

Training deep learning models for semantic occupancy prediction is challenging d

ue to factors such as a large number of occupancy cells, severe occlusion, limit ed visual cues, complicated driving scenarios, etc. Recent methods often adopt t ransformer-based architectures given their strong capability in learning input-c onditioned weights and long-range relationships. However, transformer-based netw orks are notorious for their quadratic computation complexity, seriously undermi ning their efficacy and deployment in semantic occupancy prediction. Inspired by the global modeling and linear computation complexity of the Mamba architecture , we present the first Mamba-based network for semantic occupancy prediction, te rmed OccMamba. Specifically, we first design the hierarchical Mamba module and l ocal context processor to better aggregate global and local contextual informati on, respectively. Besides, to relieve the inherent domain gap between the lingui stic and 3D domains, we present a simple yet effective 3D-to-1D reordering schem e, i.e., height-prioritized 2D Hilbert expansion. It can maximally retain the sp atial structure of 3D voxels as well as facilitate the processing of Mamba block s. Endowed with the aforementioned designs, our OccMamba is capable of directly and efficiently processing large volumes of dense scene grids, achieving state-o f-the-art performance across three prevalent occupancy prediction benchmarks, in cluding OpenOccupancy, SemanticKITTI, and SemanticPOSS. Notably, on OpenOccupanc y, our OccMamba outperforms the previous state-of-the-art Co-Occ by 5.1% IoU and 4.3% mIoU, respectively. Our implementation is open-sourced and available at: h ttps://github.com/USTCLH/OccMamba.

****

3D Convex Splatting: Radiance Field Rendering with 3D Smooth Convexes

Jan Held, Renaud Vandeghen, Abdullah Hamdi, Adrien Deliege, Anthony Cioppa, Silv io Giancola, Andrea Vedaldi, Bernard Ghanem, Marc Van Droogenbroeck; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 213 60-21369

Recent advances in radiance field reconstruction, such as 3D Gaussian Splatting (3DGS), have achieved high-quality novel view synthesis and fast rendering by re presenting scenes with compositions of Gaussian primitives. However, 3D Gaussian s present several limitations for scene reconstruction. Accurately capturing har d edges is challenging without significantly increasing the number of Gaussians, creating a large memory footprint. Moreover, they struggle to represent flat su rfaces, as they are diffused in space. Without hand-crafted regularizers, they t end to disperse irregularly around the actual surface. To circumvent these issue s, we introduce a novel method, named 3D Convex Splatting (3DCS), which leverage s 3D smooth convexes as primitives for modeling geometrically-meaningful radianc e fields from multi-view images. Smooth convex shapes offer greater flexibility than Gaussians, allowing for a better representation of 3D scenes with hard edge s and dense volumes using fewer primitives. Powered by our efficient CUDA-based rasterizer, 3DCS achieves superior performance over 3DGS on benchmarks such as M ip-NeRF360, Tanks and Temples, and Deep Blending. Specifically, our method attai ns an improvement of up to 0.81 in PSNR and 0.026 in LPIPS compared to 3DGS whil e maintaining high rendering speeds and reducing the number of required primitiv es. Our results highlight the potential of 3D Convex Splatting to become the new standard for high-quality scene reconstruction and novel view synthesis. The pr oject page is https://convexsplatting.github.io/ .

****

3D Prior Is All You Need: Cross-Task Few-shot 2D Gaze Estimation

Yihua Cheng, Hengfei Wang, Zhongqun Zhang, Yang Yue, Boeun Kim, Feng Lu, Hyung J in Chang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23891-23900

3D and 2D gaze estimation share the fundamental objective of capturing eye movem ents but are traditionally treated as two distinct research domains. In this pap er, we introduce a novel cross-task few-shot 2D gaze estimation approach, aiming to adapt a pre-trained 3D gaze estimation network for 2D gaze prediction on uns een devices using only a few training images. This task is highly challenging du e to the domain gap between 3D and 2D gaze, unknown screen poses, and limited tr aining data. To address these challenges, we propose a novel framework that brid ges the gap between 3D and 2D gaze. Our framework contains a physics-based diffe

rentiable projection module with learnable parameters to model screen poses and project 3D gaze into 2D gaze. The framework is fully differentiable and can integrate into existing 3D gaze networks without modifying their original architecture. Additionally, we introduce a dynamic pseudo-labelling strategy for flipped images, which is particularly challenging for 2D labels due to unknown screen poses. To overcome this, we reverse the projection process by converting 2D labels to 3D space, where flipping is performed. Notably, this 3D space is not aligned with the camera coordinate system, so we learn a dynamic transformation matrix to compensate for this misalignment. We evaluate our method on MPIIGaze, EVE, and GazeCapture datasets, collected respectively on laptops, desktop computers, and mobile devices. The superior performance highlights the effectiveness of our approach, and demonstrates its strong potential for real-world applications.
****

Cheb-GR: Rethinking K-nearest Neighbor Search in Re-ranking for Person Re-identification

Jinxi Yang, He Li, Bo Du, Mang Ye; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19261-19270

Person re-identification (ReID) is the task of matching individuals across different camera views. Existing approaches typically employ neural networks to extract discriminative features, ranking gallery images based on their similarities to probe images. While effective, these methods are often enhanced through re-ranking, a post-processing step that refines initial retrieval results without requiring additional model training. However, current re-ranking methods mostly rely on k-nearest neighbor search to extract similar images that might have the same identity as the query, which is time-consuming with a high computation burden, limiting their applications in reality. We rethink the effect of the k-nearest neighbor search and introduce the Chebyshev's Theorem-guided Graph Re-ranking (Cheb-GR) method, which adopts the adaptive neighbor search guided by Chebyshev's Theorem over the k-nearest neighbor search for efficient neighbor selection. Our method leverages graph convolution operations to refine image features and achieve robust re-ranking, leading to enhanced retrieval performance. Furthermore, we provide a theoretical analysis based on Chebyshev's Inequality to elucidate the factors contributing to the strong performance of the proposed method. Our method significantly reduces the computation costs while maintaining relatively strong performance. Through extensive experiments in both general and cross-domain settings, we demonstrate the effectiveness of Cheb-GR and its potential for real-world applications.
****

Spotting the Unexpected (STU): A 3D LiDAR Dataset for Anomaly Segmentation in Autonomous Driving

Alexey Nekrasov, Malcolm Burdorf, Stewart Worrall, Bastian Leibe, Julie Stephany Berrio Perez; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11875-11885

To operate safely, autonomous vehicles (AVs) need to detect and handle unexpected objects or anomalies on the road. While significant research exists for anomaly detection and segmentation in 2D, research progress in 3D is underexplored. Existing datasets lack high-quality multimodal data that are typically found in AVs. This paper presents a novel dataset for anomaly segmentation in driving scenarios. To the best of our knowledge, it is the first publicly available dataset focused on road anomaly segmentation with dense 3D semantic labeling, incorporating both LiDAR and camera data, as well as sequential information to enable anomaly detection across various ranges. This capability is critical for the safe navigation of autonomous vehicles. We adapted and evaluated several baseline models for 3D segmentation, highlighting the challenges of 3D anomaly detection in driving environments. Our dataset and evaluation code will be openly available, facilitating the testing and performance comparison of different approaches.
****

Is `Right' Right? Enhancing Object Orientation Understanding in Multimodal Large Language Models through Egocentric Instruction Tuning

Ji Hyeok Jung, Eun Tae Kim, Seoyeon Kim, Joo Ho Lee, Bumsoo Kim, Buru Chang; Pro

ceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14257-14267
Multimodal large language models (MLLMs) act as essential interfaces, connecting humans with AI technologies in multimodal applications. However, current MLLMs face challenges in accurately interpreting object orientation in images due to inconsistent orientation annotations in training data, hindering the development of a coherent orientation understanding. To overcome this, we propose egocentric instruction tuning, which aligns MLLMs' orientation understanding with the user's perspective, based on a consistent annotation standard derived from the user's egocentric viewpoint. We first generate egocentric instruction data that leverages MLLMs' ability to recognize object details and applies prior knowledge for orientation understanding. Using this data, we perform instruction tuning to enhance the model's capability for accurate orientation interpretation. In addition, we introduce EgoOrientBench, a benchmark that evaluates MLLMs' orientation understanding across three tasks using images collected from diverse domains. Experimental results on this benchmark show that egocentric instruction tuning significantly improves orientation understanding without compromising overall MLLM performance. The instruction data and benchmark dataset are available on our project page at https://github.com/jhCOR/EgoOrientBench.
****

GCC: Generative Color Constancy via Diffusing a Color Checker
Chen-Wei Chang, Cheng-De Fan, Chia-Che Chang, Yi-Chen Lo, Yu-Chee Tseng, Jiun-Long Huang, Yu-Lun Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10868-10878
Color constancy methods often struggle to generalize across different camera sensors due to varying spectral sensitivities. We present GCC, which leverages diffusion models to inpaint color checkers into images for illumination estimation. Our key innovations include (1) a single-step deterministic inference approach that inpaints color checkers reflecting scene illumination, (2) a Laplacian decomposition technique that preserves checker structure while allowing illumination-dependent color adaptation, and (3) a mask-based data augmentation strategy for handling imprecise color checker annotations. By harnessing rich priors from pre-trained diffusion models, GCC demonstrates strong robustness in challenging cross-camera scenarios. These results highlight our method's effective generalization capability across different camera characteristics without requiring sensor-specific training, making it a versatile and practical solution for real-world applications.
****

Do Visual Imaginations Improve Vision-and-Language Navigation Agents?
Akhil Perincherry, Jacob Krantz, Stefan Lee; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3846-3855
Vision-and-Language Navigation (VLN) agents are tasked with navigating an unseen environment using natural language instructions. In this work, we study if visual representations of sub-goals implied by the instructions can serve as navigational cues and lead to increased navigation performance. To synthesize these visual representations or "imaginations", we leverage a text-to-image diffusion model on landmark references contained in segmented instructions. These imaginations are provided to VLN agents as an added modality to act as landmark cues and an auxiliary loss is added to explicitly encourage relating these with their corresponding referring expressions. Our findings reveal an increase in success rate (SR) of ~1 point and up to ~0.5 points in success scaled by inverse path length (SPL) across agents. These results suggest that the proposed approach reinforces visual understanding compared to relying on language instructions alone.
****

On Denoising Walking Videos for Gait Recognition
Dongyang Jin, Chao Fan, Jingzhe Ma, Jingkai Zhou, Weihua Chen, Shiqi Yu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12347-12357
To capture individual gait patterns, excluding identity-irrelevant cues in walking videos, such as clothing texture and color, remains a persistent challenge fo

r vision-based gait recognition. Traditional silhouette and pose-based methods, though theoretically effective at removing such distractions, often fall short of high accuracy due to their sparse and less informative inputs. To address this, emerging end-to-end methods focus on directly denoising RGB videos using global optimization and human-defined priors. Building on this trend, we propose a novel gait denoising method, DenosingGait. Inspired by the philosophy that "what I cannot create, I do not understand", we turn to generative diffusion models, uncovering how these models can partially filter out irrelevant factors for improved gait understanding. Based on this generation-driven denoising, we introduce feature matching, a kind of popular geometrical constraint in optical flow and depth estimation, to compact multi-channel float-encoded RGB information into two-channel direction vectors that represent local structural features, where within-frame matching captures spatial details and cross-frame matching conveys temporal dynamics. Experiments on the CCPG, CAISA-B*, and SUSTech1K datasets demonstrate that DenoisingGait achieves a new SoTA performance in most cases for both within-domain and cross-domain evaluations. Code is available at https://github.com/ShiqiYu/OpenGait.
****

Conformal Prediction for Zero-Shot Models
Julio Silva-Rodríguez, Ismail Ben Ayed, Jose Dolz; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19931-19941
Vision-language models pre-trained at large scale have shown unprecedented adaptability and generalization to downstream tasks. Although its discriminative potential has been widely explored, its reliability and uncertainty are still overlooked. In this work, we investigate the capabilities of CLIP models under the split conformal prediction paradigm, which provides theoretical guarantees to black-box models based on a small, labeled calibration set. In contrast to the main body of literature on conformal predictors in vision classifiers, foundation models exhibit a particular characteristic: they are pre-trained on a one-time basis on an inaccessible source domain, different from the transferred task. This domain drift negatively affects the efficiency of the conformal sets and poses additional challenges. To alleviate this issue, we propose Conf-OT, a transfer learning setting that operates transductive over the combined calibration and query sets. Solving an optimal transport problem, the proposed method bridges the domain gap between pre-training and adaptation without requiring additional data splits but still maintaining coverage guarantees. We comprehensively explore this conformal prediction strategy on a broad span of 15 datasets and three non-conformity scores. Conf-OT provides consistent relative improvements of up to 20% on set efficiency while being 15 times faster than popular transductive approaches.
****

PhysAnimator: Physics-Guided Generative Cartoon Animation
Tianyi Xie, Yiwei Zhao, Ying Jiang, Chenfanfu Jiang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10793-10804
Creating hand-drawn animation sequences is labor-intensive and demands professional expertise. We introduce PhysAnimator, a novel approach for generating physically plausible meanwhile anime-stylized animation from static anime illustrations. Our method seamlessly integrates physics-based simulations with data-driven generative models to produce dynamic and visually compelling animations. To capture the fluidity and exaggeration characteristic of anime, we perform image-space deformable body simulations on extracted mesh geometries. We enhance artistic control by introducing customizable energy strokes and incorporating rigging point support, enabling the creation of tailored animation effects such as wind interactions. Finally, we extract and warp sketches from the simulation sequence, generating a texture-agnostic representation, and employ a sketch-guided video diffusion model to synthesize high-quality animation frames. The resulting animations exhibit temporal consistency and visual plausibility, demonstrating the effectiveness of our method in creating dynamic anime-style animations.
****

SeriesBench: A Benchmark for Narrative-Driven Drama Series Understanding
Chenkai Zhang, Yiming Lei, Zeming Liu, Haitao Leng, ShaoGuo Liu, Tingting Gao, Q

ingjie Liu, Yunhong Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28995-29004
With the rapid development of Multi-modal Large Language Models (MLLMs), an increasing number of benchmarks have been established to evaluate the video understanding capabilities of these models. However, these benchmarks focus on standalone videos and only assess "visual elements" like human actions and object states. In reality, contemporary videos often encompass complex and continuous narratives, typically presented as a series. To address this challenge, we propose SeriesBench, a benchmark consisting of 105 carefully curated narrative-driven series, covering 28 specialized tasks that require deep narrative understanding to solve. Specifically, we first select a diverse set of drama series spanning various genres. Then, we introduce a novel long-span narrative annotation method, combined with a full-information transformation approach to convert manual annotations into diverse task formats. To further enhance the model's capacity for detailed analysis of plot structures and character relationships within series, we propose a novel narrative reasoning framework, PC-DCoT. Extensive results on SeriesBench indicate that existing MLLMs still face significant challenges in understanding narrative-driven series, while PC-DCoT enables these MLLMs to achieve performance improvements. Overall, our SeriesBench and PC-DCoT highlight the critical necessity of advancing model capabilities for understanding narrative-driven series, guiding future MLLMs development. SeriesBench is publicly available at https://github.com/zackhxn/SeriesBench-CVPR2025.
****

Weakly Supervised Temporal Action Localization via Dual-Prior Collaborative Learning Guided by Multimodal Large Language Models
Quan Zhang, Jinwei Fang, Rui Yuan, Xi Tang, Yuxin Qi, Ke Zhang, Chun Yuan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24139-24148
Recent breakthroughs in Multimodal Large Language Models (MLLMs) have gained significant recognition within the deep learning community, where the fusion of the Video Foundation Models (VFMs) and Large Language Models(LLMs) has proven instrumental in constructing robust video understanding systems, effectively surmounting constraints associated with predefined visual tasks. These sophisticated MLLMs exhibit remarkable proficiency in comprehending videos, swiftly attaining unprecedented performance levels across diverse benchmarks. However, their operation demands substantial memory and computational resources, underscoring the continued importance of traditional models in video comprehension tasks. In this paper, we introduce a novel learning paradigm termed MLLM4WTAL. This paradigm harnesses the potential of MLLM to offer temporal action key semantics and complete semantic textual cues for conventional Weakly-supervised Temporal Action Localization (WTAL) methods. MLLM4WTAL facilitates the enhancement of WTAL by leveraging MLLM guidance. It achieves this by integrating two distinct modules: Key Semantic Matching (KSM) and Complete Semantic Reconstruction (CSR). These modules work in tandem to effectively address prevalent issues like incomplete and over-complete outcomes common in WTAL methods. Rigorous experiments are conducted to validate the efficacy of our proposed approach in augmenting the performance of various heterogeneous WTAL models.
****

FIMA-Q: Post-Training Quantization for Vision Transformers by Fisher Information Matrix Approximation
Zhuguanyu Wu, Shihe Wang, Jiayi Zhang, Jiaxin Chen, Yunhong Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14891-14900
Post-training quantization (PTQ) has stood out as a cost-effective and promising model compression approach over recent years, as it eliminates the need for retraining on the entire dataset. Unfortunately, most existing PTQ methods for Vision Transformers (ViTs) exhibit a notable drop in accuracy, especially in low-bit cases. To tackle these challenges, we analyze the extensively utilized Hessian-guided quantization loss, and uncover certain limitations within the approximated pre-activation Hessian. Following the block-by-block reconstruction paradigm o

f PTQ, we first derive a quantization loss based on the Fisher Information Matrix (FIM). Due to the large scale of the complete FIM, we establish the relationship between KL divergence and FIM in the PTQ scenario to enable fast computation of the quantization loss during reconstruction. Subsequently, we develop a Diagonal Plus Low-Rank (DPLR) estimation on FIM to achieve a more nuanced quantization loss. Our extensive experiments, conducted across various vision tasks with distinct representative ViT-based architectures on public benchmark datasets, demonstrate that our method outperforms the state-of-the-art approaches, especially in the case of low-bit quantization. The source code is available at https://github.com/ShiheWang/FIMA-Q.

****

HotSpot: Signed Distance Function Optimization with an Asymptotically Sufficient Condition

Zimo Wang, Cheng Wang, Taiki Yoshino, Sirui Tao, Ziyang Fu, Tzu-Mao Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1276-1286

We propose a method, HotSpot, for optimizing neural signed distance functions. Existing losses, such as the eikonal loss, act as necessary but insufficient constraints and cannot guarantee that the recovered implicit function represents a true distance function, even if the output minimizes these losses almost everywhere. Furthermore, the eikonal loss suffers from stability issues in optimization. Finally, in conventional methods, regularization losses that penalize surface area distort the reconstructed signed distance function. We address these challenges by designing a loss function using the solution of a screened Poisson equation. Our loss, when minimized, provides an asymptotically sufficient condition to ensure the output converges to a true distance function. Our loss also leads to stable optimization and naturally penalizes large surface areas. We present theoretical analysis and experiments on both challenging 2D and 3D datasets and show that our method provides better surface reconstruction and a more accurate distance approximation.

****

GliaNet: Adaptive Neural Network Structure Learning with Glia-Driven

Mengqiao Han, Liyuan Pan, Xiabi Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25240-25249

Neural networks derived from the M-P model have excelled in various visual tasks. However, as a simplified simulation version of the brain neural pathway, their structures are locked during training, causing over-fitting and over-parameterization. Although recent models have begun using the biomimetic concept and empirical pruning, they still result in irrational pruning, potentially affecting the accuracy of the model. In this paper, we introduce the Glia unit, composed of oligodendrocytes (Oli) and astrocytes (Ast), to emulate the exact workflow of the mammalian brain, thereby enhancing the biological plausibility of neural functions. Oli selects neurons involved in signal transmission during neural communication and, together with Ast, adaptively optimizes the neural structure. Specifically, we first construct the artificial Glia-Neuron (G-N) model, which is formulated at the instance, group, and interaction levels with adaptive and collaborative mechanisms. Then, we construct GliaNet based on our G-N model, whose structure and connections can be continuously optimized during training. Experiments show that our GliaNet advances state-of-the-art on multiple tasks while significantly reducing its parameters.

****

BACON: Improving Clarity of Image Captions via Bag-of-Concept Graphs

Zhantao Yang, Ruili Feng, Keyu Yan, Huangji Wang, Zhicai Wang, Shangwen Zhu, Han Zhang, Jie Xiao, Pingyu Wu, Kai Zhu, Jixuan Chen, Chen-Wei Xie, Yue Yang, Hongyang Zhang, Yu Liu, Fan Cheng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14380-14389

Advancements in large Vision-Language Models have brought precise, accurate image captioning, vital for advancing multi-modal image understanding and processing. Yet these captions often carry lengthy, intertwined contexts that are difficult to parse and frequently overlook essential cues, posing a great barrier for mo

dels like GroundingDINO and SDXL, which lack the strong text encoding and syntax analysis needed to fully leverage dense captions.To address this, we propose BACON, a prompting method that breaks down VLM-generated captions into disentangled, structured elements such as objects, relationships, styles, and themes. This approach not only minimizes confusion from handling complex contexts but also allows for efficient transfer into a JSON dictionary, enabling models without linguistic processing capabilities to easily access key information.We annotated 100,000 image-caption pairs using BACON with GPT-4V and trained an LLaVA captioner on this dataset, enabling it to produce BACON-style captions without relying on costly GPT-4V resources. Evaluations of overall quality, precision, and recall--as well as user studies--demonstrate that the resulting caption model consistently outperforms other state-of-the-art VLM models in generating high-quality captions.Additionally, we show that BACON-style captions exhibit better clarity when applied to various models, enabling them to accomplish previously unattainable tasks or surpass existing SOTA solutions without training. For example, BACON-style captions help groundingDINO achieve 1.51 times higher recall scores on open-vocabulary object detection tasks compared to leading methods.
****

EntitySAM: Segment Everything in Video
Mingqiao Ye, Seoung Wug Oh, Lei Ke, Joon-Young Lee; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24234-24243
Automatically tracking and segmenting every video entity remains a significant challenge. Despite rapid advancements in video segmentation, even state-of-the-art models like SAM 2 struggle to consistently track all entities across a video--a task we refer to as Video Entity Segmentation.We propose EntitySAM, a framework for zero-shot video entity segmentation. EntitySAM extends SAM 2 by removing the need for explicit prompts, allowing automatic discovery and tracking of all entities, including those appearing in later frames. We incorporate query-based entity discovery and association into SAM 2, inspired by transformer-based object detectors. Specifically, we introduce an entity decoder to facilitate inter-object communication and an automatic prompt generator using learnable object queries. Additionally, we add a semantic encoder to enhance SAM 2's semantic awareness, improving segmentation quality. Trained on image-level mask annotations without category information from the COCO dataset, EntitySAM demonstrates strong generalization on four zero-shot video segmentation tasks: Video Entity, Panoptic, Instance, and Semantic Segmentation. Results on six popular benchmarks show that EntitySAM outperforms previous unified video segmentation methods and strong baselines, setting new standards for zero-shot video segmentation.
****

GS-2DGS: Geometrically Supervised 2DGS for Reflective Object Reconstruction
Jinguang Tong, Xuesong Li, Fahira Afzal Maken, Sundaram Muthu, Lars Petersson, Chuong Nguyen, Hongdong Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21547-21557
3D modeling of highly reflective objects remains challenging due to strong view-dependent appearances. While previous SDF-based methods can recover high-quality meshes, they are often time-consuming and tend to produce over-smoothed surfaces. In contrast, 3D Gaussian Splatting (3DGS) offers the advantage of high speed and detailed real-time rendering, but extracting surfaces from the Gaussians can be noisy due to the lack of geometric constraints. To bridge the gap between these approaches, we propose a novel reconstruction method called GS-2DGS for reflective objects based on 2D Gaussian Splatting (2DGS). Our approach combines the rapid rendering capabilities of Gaussian Splatting with additional geometric information from a foundation model. Experimental results on synthetic and real datasets demonstrate that our method significantly outperforms Gaussian-based techniques in terms of reconstruction and relighting and achieves performance comparable to SDF-based methods while being an order of magnitude faster.
****

Libra-Merging: Importance-redundancy and Pruning-merging Trade-off for Acceleration Plug-in in Large Vision-Language Model
Longrong Yang, Dong Shen, Chaoxiang Cai, Kaibing Chen, Fan Yang, Tingting Gao, D

i Zhang, Xi Li; Proceedings of the Computer Vision and Pattern Recognition Confe rence (CVPR), 2025, pp. 9402-9412
Large Vision-Language Models (LVLMs) have achieved significant progress in recen t years. However, the expensive inference cost limits the realistic deployment o f LVLMs. Some works find that visual tokens are redundant and compress tokens to reduce the inference cost. These works identify important non-redundant tokens as target tokens, then prune the remaining tokens (non-target tokens) or merge t hem into target tokens. However, target token identification faces the token imp ortance-redundancy dilemma. Besides, token merging and pruning face a dilemma be tween disrupting target token information and losing non-target token informatio n. To solve these problems, we propose a novel visual token compression scheme, named Libra-Merging. In target token identification, Libra-Merging selects the m ost important tokens from spatially discrete intervals, achieving a more robust token importance-redundancy trade-off than relying on a hyper-parameter. In toke n compression, when non-target tokens are dissimilar to target tokens, Libra-Mer ging does not merge them into the target tokens, thus avoiding disrupting target token information. Meanwhile, Libra-Merging condenses these non-target tokens i nto an information compensation token to prevent losing important non-target tok en information. Our method can serve as a plug-in for diverse LVLMs, and extensi ve experimental results demonstrate its effectiveness. The code will be publicly available at https://github.com/longrongyang/Libra-Merging.
****

VasTSD: Learning 3D Vascular Tree-state Space Diffusion Model for Angiography Sy nthesis
Zhifeng Wang, Renjiao Yi, Xin Wen, Chenyang Zhu, Kai Xu; Proceedings of the Comp uter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15693-15702
Angiography imaging is a medical imaging technique that enhances the visibility of blood vessels within the body by using contrast agents. Angiographic images c an effectively assist in the diagnosis of vascular diseases. However, contrast a gents may bring extra radiation exposure which is harmful to patients with healt h risks. To mitigate these concerns, in this paper, we aim to automatically gene rate angiography from non-angiographic inputs, by leveraging and enhancing the i nherent physical properties of vascular structures. Previous methods relying on 2D slice-based angiography synthesis struggle with maintaining continuity in 3D vascular structures and exhibit limited effectiveness across different imaging m odalities. We propose VasTSD, a 3D vascular tree-state space diffusion model to synthesize angiography from 3D non-angiographic volumes, with a novel state spac e serialization approach that dynamically constructs vascular tree topologies, i ntegrating these with a diffusion-based generative model to ensure the generatio n of anatomically continuous vasculature in 3D volumes. A pre-trained vision emb edder is employed to construct vascular state space representations, enabling co nsistent modeling of vascular structures across multiple modalities. Extensive e xperiments on various angiographic datasets demonstrate the superiority of VasTS D over prior works, achieving enhanced continuity of blood vessels in synthesize d angiographic synthesis in multiple modalities and anatomical regions.
****

PanSplat: 4K Panorama Synthesis with Feed-Forward Gaussian Splatting
Cheng Zhang, Haofei Xu, Qianyi Wu, Camilo Cruz Gambardella, Dinh Phung, Jianfei Cai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11437-11447
With the advent of portable 360deg cameras, panorama has gained significant atte ntion in applications like virtual reality (VR), virtual tours, robotics, and au tonomous driving. As a result, wide-baseline panorama view synthesis has emerged as a vital task, where high resolution, fast inference, and memory efficiency a re essential. Nevertheless, existing methods typically focus on lower resolution s (512 x1024) due to demanding memory and computational requirements. In this pa per, we present PanSplat, a generalizable, feed-forward approach that efficientl y supports resolution up to 4K (2048 x4096). Our approach features a tailored sp herical 3D Gaussian pyramid with a Fibonacci lattice arrangement, enhancing imag e quality while reducing information redundancy. To accommodate the demands of h

igh resolution, we propose a pipeline that integrates a hierarchical spherical c ost volume and localized Gaussian heads, enabling two-step deferred backpropagat ion for memory-efficient training on a single A100 GPU. Experiments demonstrate that PanSplat achieves state-of-the-art results with superior efficiency and ima ge quality across both synthetic and real-world datasets.
****

WISNet: Pseudo Label Generation on Unbalanced and Patch Annotated Waste Images
Shifan Zhang, Hongzi Zhu, Yinan He, Minyi Guo, Ziyang Lou, Shan Chang; Proceedin gs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1 5076-15085
Computer-vision-based assessment on waste sorting is desired to replace manpower supervision in Shanghai city. Due to the hardness of labeling a multitude of wa ste images, it is infeasible to train a semantic segmentation model for this pur pose directly. In this work, we construct a new dataset consisting of 12,208 was te images, upon which seed regions (i.e., patches) are annotated and classified into 21 categories in a crowdsourcing fashion. To obtain pixel-level labels to t rain an effective segmentation model, we propose a weakly-supervised waste image pseudo label generation scheme, called WISNet. Specifically, we train a cohesiv e feature extractor with contrastive prototype learning, incorporating an unsupe rvised classification pretext task to help the extractor focus on more discrimin ative regions even with the same category. Furthermore, we propose an effective iterative patch expansion method to generate accurate pixel-level pseudo labels. Given these generated pseudo labels, a few-shot segmentation model can be train ed to segment waste images. We implement and deploy WISNet in two real-world sce narios and conduct intensive experiments. Results show that WISNet can achieve a state-of-the-art 40.2% final segmentation mIoU on our waste benchmark, outperfo rming all other baselines and demonstrating the efficacy of WISNet.
****

MixerMDM: Learnable Composition of Human Motion Diffusion Models
Pablo Ruiz-Ponce, German Barquero, Cristina Palmero, Sergio Escalera, José Garcí a-Rodríguez; Proceedings of the Computer Vision and Pattern Recognition Conferen ce (CVPR), 2025, pp. 12380-12390
Generating human motion guided by conditions such as textual descriptions is cha llenging due to the need for datasets with pairs of high-quality motion and thei r corresponding conditions. The difficulty increases when aiming for finer contr ol in the generation. To that end, prior works have proposed to combine several motion diffusion models pre-trained on datasets with different types of conditio ns, thus allowing control with multiple conditions. However, the proposed mergin g strategies overlook that the optimal way to combine the generation processes m ight depend on the particularities of each pre-trained generative model and also the specific textual descriptions. In this context, we introduce MixerMDM, the first learnable model composition technique for combining pre-trained text-condi tioned human motion diffusion models. Unlike previous approaches, MixerMDM provi des a dynamic mixing strategy that is trained in an adversarial fashion to learn to combine the denoising process of each model depending on the set of conditio ns driving the generation. By using MixerMDM to combine single- and multi-person motion diffusion models, we achieve fine-grained control on the dynamics of eve ry person individually, and also on the overall interaction. Furthermore, we pro pose a new evaluation technique that, for the first time in this task, measures the interaction and individual quality by computing the alignment between the mi xed generated motions and their conditions as well as the capabilities of MixerM DM to adapt the mixing throughout the denoising process depending on the motions to mix.
****

Hand-held Object Reconstruction from RGB Video with Dynamic Interaction
Shijian Jiang, Qi Ye, Rengan Xie, Yuchi Huo, Jiming Chen; Proceedings of the Com puter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12220-12230
This work aims to reconstruct the 3D geometry of a rigid object manipulated by o ne or both hands using monocular RGB video. Previous methods rely on Structure-f rom-Motion or hand priors to estimate relative motion between the object and cam

era, which typically assume textured objects or single-hand interactions. To acc
urately recover object geometry in dynamic interactions, we incorporate priors f
rom 3D generation model into object pose estimation and propose semantic consist
ency constraints to solve the challenge of shape and texture discrepancy between
 the generated priors and observations. The poses are initialized, followed by j
oint optimization of the object poses and implicit neural representation. During
 optimization, a novel pose outlier voting strategy with inter-view consistency
is proposed to correct large pose errors. Experiments on three datasets demonstr
ate that our method significantly outperforms the state-of-the-art in reconstruc
tion quality for both single- and two-hand scenarios. Our project page: https://
east-j.github.io/dynhor/
****

LEDiff: Latent Exposure Diffusion for HDR Generation
Chao Wang, Zhihao Xia, Thomas Leimkuhler, Karol Myszkowski, Xuaner Zhang; Procee
dings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp
. 453-464
While consumer displays increasingly support more than 10 stops of dynamic range
, most image assets -- such as internet photographs and generative AI content --
 remain limited to 8-bit low dynamic range (LDR), constraining their utility acr
oss high dynamic range (HDR) applications. Currently, no generative model can pr
oduce high-bit, high-dynamic range content in a generalizable way. Existing LDR-
to-HDR conversion methods often struggle to produce photorealistic details and p
hysically-plausible dynamic range in the clipped areas. We introduce LEDiff, a m
ethod that enables a generative model with HDR content generation through latent
 space fusion inspired by image-space exposure fusion techniques. It also functi
ons as an LDR-to-HDR converter, expanding the dynamic range of existing low-dyna
mic range images. Our approach uses a small HDR dataset to enable a pretrained d
iffusion model to recover detail and dynamic range in clipped highlights and sha
dows. LEDiff brings HDR capabilities to existing generative models and converts
any LDR image to HDR, creating photorealistic HDR outputs for image generation,
image-based lighting (HDR environment map generation), and photographic effects
such as depth of field simulation, where linear HDR data is essential for realis
tic quality.
****

Video Depth Anything: Consistent Depth Estimation for Super-Long Videos
Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, Bi
ngyi Kang; Proceedings of the Computer Vision and Pattern Recognition Conference
 (CVPR), 2025, pp. 22831-22840
Depth Anything has achieved remarkable success in monocular depth estimation wit
h strong generalization ability. However, it suffers from temporal inconsistency
 in videos, hindering its practical applications. Various methods have been prop
osed to alleviate this issue by leveraging video generation models or introducin
g priors from optical flow and camera poses. Nonetheless, these methods are only
 applicable to short videos (10 seconds) and require a trade-off between quality
 and computational efficiency. We propose Video Depth Anything for high-quality,
 consistent depth estimation in super-long videos (over several minutes) without
 sacrificing efficiency. We base our model on Depth Anything V2 and replace its
head with an efficient spatial-temporal head. We design a straightforward yet ef
fective temporal consistency loss by constraining the temporal depth gradient, e
liminating the need for additional geometric priors. The model is trained on a j
oint dataset of video depth and unlabeled images, similar to Depth Anything V2.
Moreover, a novel key-frame-based strategy is developed for long video inference
. Experiments show that our model can be applied to arbitrarily long videos wit
hout compromising quality, consistency, or generalization ability. Comprehensive
 evaluations on multiple video benchmarks demonstrate that our approach sets a n
ew state-of-the-art in zero-shot video depth estimation. We offer models of diff
erent scales to support a range of scenarios, with our smallest model capable of
 real-time performance at 30 FPS.
****

VideoAutoArena: An Automated Arena for Evaluating Large Multimodal Models in Vid

eo Analysis through User Simulation
Ziyang Luo, Haoning Wu, Dongxu Li, Jing Ma, Mohan Kankanhalli, Junnan Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8461-8474
Large multimodal models (LMMs) with advanced video analysis capabilities have recently garnered significant attention. However, most evaluations rely on traditional methods like multiple-choice question answering in benchmarks such as VideoMME and LongVideoBench, which are prone to lack the depth needed to capture the complex demands of real-world users. To address this limitation--and due to the prohibitive cost and slow pace of human annotation for video tasks--we introduce VideoAutoArena, an arena-style benchmark inspired by LMSYS Chatbot Arena's framework, designed to automatically assess LMMs' video analysis abilities. VideoAutoArena utilizes user simulation to generate open-ended, adaptive questions that rigorously assess model performance in video understanding. The benchmark features an automated, scalable evaluation framework, incorporating a modified ELO Rating System for fair and continuous comparisons across multiple LMMs. To validate our automated judging system, we construct a "gold standard" using a carefully curated subset of human annotations, demonstrating that our arena strongly aligns with human judgment while maintaining scalability. Additionally, we introduce a fault-driven evolution strategy, progressively increasing question complexity to push models toward handling more challenging video analysis scenarios. Experimental results demonstrate that VideoAutoArena effectively differentiates among state-of-the-art LMMs, providing insights into model strengths and areas for improvement. To further streamline our evaluation, we introduce VideoAutoBench as an auxiliary benchmark, where human annotators label winners in a subset of VideoAutoArena battles. We use GPT-4o as a judge to compare responses against these human-validated answers. Together, VideoAutoArena and VideoAutoBench offer a cost-effective, and scalable framework for evaluating LMMs in user-centric video analysis.
****

InstanceCap: Improving Text-to-Video Generation via Instance-aware Structured Caption
Tiehan Fan, Kepan Nan, Rui Xie, Penghao Zhou, Zhenheng Yang, Chaoyou Fu, Xiang Li, Jian Yang, Ying Tai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28974-28983
Text-to-video generation has evolved rapidly in recent years, delivering remarkable results. Training typically relies on video-caption paired data, which plays a crucial role in enhancing generation performance. However, current video captions often suffer from insufficient details, hallucinations and imprecise motion depiction, affecting the fidelity and consistency of generated videos. In this work, we propose a novel instance-aware structured caption framework, termed \mathtt InstanceCap , to achieve instance-level and fine-grained video caption for the first time. Based on this scheme, we design an auxiliary models cluster to convert original video into instances to enhance instance fidelity. Video instances are further used to refine dense prompts into structured phrases, achieving concise yet precise descriptions. Furthermore, a 22K \mathtt InstanceVid dataset is curated for training, and an enhancement pipeline that tailored to \mathtt InstanceCap structure is proposed for inference. Experimental results demonstrate that our proposed \mathtt InstanceCap significantly outperform previous models, ensuring high fidelity between captions and videos while reducing hallucinations.
****

AudCast: Audio-Driven Human Video Generation by Cascaded Diffusion Transformers
Jiazhi Guan, Kaisiyuan Wang, Zhiliang Xu, Quanwei Yang, Yasheng Sun, Shengyi He, Borong Liang, Yukang Cao, Yingying Li, Haocheng Feng, Errui Ding, Jingdong Wang, Youjian Zhao, Hang Zhou, Ziwei Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10678-10689
Despite the recent progress of audio-driven video generation, existing methods mostly focus on driving facial movements, leading to non-coherent head and body dynamics. Moving forward, it is desirable yet challenging to generate holistic hu

man videos with both accurate lip-sync and delicate co-speech gestures w.r.t. gi ven audio. In this work, we propose AudCast, a generalized audio-driven human vi deo generation framework adopting a cascade Diffusion-Transformers (DiTs) paradi gm, which synthesizes holistic human videos based on a reference image and a giv en audio. 1) Firstly, an audio-conditioned Holistic Human DiT architecture is pr oposed to directly drive the movements of any human body with vivid gesture dyna mics. 2) Then to enhance hand and face details that are well-knownly difficult t o handle, a Regional Refinement DiT leverages regional 3D fitting as the bridge to reform the signals, producing the final results. Extensive experiments demon strate that our framework generates high-fidelity audio-driven holistic human vi deos with temporal coherence and fine facial and hand details.
****

Luminance-GS: Adapting 3D Gaussian Splatting to Challenging Lighting Conditions with View-Adaptive Curve Adjustment

Ziteng Cui, Xuangeng Chu, Tatsuya Harada; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26472-26482

Capturing high-quality photographs under diverse real-world lighting conditions is challenging, as both natural lighting (e.g., low-light) and camera exposure s ettings (e.g., exposure time) significantly impact image quality. This challenge becomes more pronounced in multi-view scenarios, where variations in lighting a nd image signal processor (ISP) settings across viewpoints introduce photometric inconsistencies. Such lighting degradations and view-dependent variations pose substantial challenges to novel view synthesis (NVS) frameworks based on Neural Radiance Fields (NeRF) and 3D Gaussian Splatting (3DGS). To address this, we int roduce Luminance-GS, a novel approach to achieving high-quality novel view synth esis results under diverse challenging lighting conditions using 3DGS. By adopti ng per-view color matrix mapping and view adaptive curve adjustments, Luminance- GS achieves state-of-the-art (SOTA) results across various lighting conditions-- including low-light, overexposure, and varying exposure--while not altering the original 3DGS explicit representation. Compared to previous NeRF- and 3DGS-based baselines, Luminance-GS provides real-time rendering speed with improved recons truction quality. The source code is available at https://github.com/cuiziteng/L uminance-GS.
****

EventSplat: 3D Gaussian Splatting from Moving Event Cameras for Real-time Render ing

Toshiya Yura, Ashkan Mirzaei, Igor Gilitschenski; Proceedings of the Computer Vi sion and Pattern Recognition Conference (CVPR), 2025, pp. 26876-26886

We introduce a method for using event camera data in novel view synthesis via Ga ussian Splatting. Event cameras offer exceptional temporal resolution and a high dynamic range. Leveraging these capabilities allows us to effectively address t he novel view synthesis challenge in the presence of fast camera motion. For ini tialization of the optimization process, our approach uses prior knowledge encod ed in an event-to-video model. We also use spline interpolation for obtaining hi gh quality poses along the event camera trajectory. This enhances the reconstruc tion quality from fast-moving cameras while overcoming the computational limitat ions traditionally associated with event-based Neural Radiance Field (NeRF) meth ods. Our experimental evaluation demonstrates that our results achieve higher vi sual fidelity and better performance than existing event-based NeRF approaches w hile being an order of magnitude faster to render.
****

Thinking in Space: How Multimodal Large Language Models See, Remember, and Recal l Spaces

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, Saining Xie; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 20 25, pp. 10632-10643

Humans possess the visual-spatial intelligence to remember spaces from sequentia l visual observations. However, can Multimodal Large Language Models (MLLMs) tra ined on million-scale video datasets also "think in space" from videos? We prese nt a novel video-based visual-spatial intelligence benchmark (VSI-Bench) of over

5,000 question-answer pairs, and find that MLLMs exhibit competitive--though subhuman--visual-spatial intelligence. We probe models to express how they think in space both linguistically and visually and find that while spatial reasoning capabilities remain the primary bottleneck for MLLMs to reach higher benchmark performance, local world models and spatial awareness do emerge within these models. Notably, prevailing linguistic reasoning techniques (e.g., chain-of-thought, self-consistency, tree-of-thoughts) fail to improve performance, whereas explicitly generating cognitive maps during question-answering enhances MLLMs' spatial distance awareness.

****

3D Student Splatting and Scooping

Jialin Zhu, Jiangbei Yue, Feixiang He, He Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21045-21054

Recently, 3D Gaussian Splatting (3DGS) provides a new framework for novel view synthesis, and has spiked a new wave of research in neural rendering and related applications. As 3DGS is becoming a foundational component of many models, any improvement on 3DGS itself can bring huge benefits. To this end, we aim to improve the fundamental paradigm and formulation of 3DGS. We argue that as an unnormalized mixture model, it needs to be neither Gaussians nor splatting. We subsequently propose a new mixture model consisting of flexible Student's t distributions, with both positive (splatting) and negative (scooping) densities. We name our model Student Splatting and Scooping, or SSS. When providing better expressivity, SSS also poses new challenges in learning. Therefore, we also propose a new principled sampling approach for optimization. Through exhaustive evaluation and comparison, across multiple datasets, settings, and metrics, we demonstrate that SSS outperforms existing methods in terms of quality and parameter efficiency, e.g. achieving matching or better quality with similar numbers of components, and obtaining comparable results while reducing the component number by as much as 82%.

****

World-consistent Video Diffusion with Explicit 3D Modeling

Qihang Zhang, Shuangfei Zhai, Miguel Ángel Bautista Martin, Kevin Miao, Alexander Toshev, Joshua Susskind, Jiatao Gu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21685-21695

Recent advancements in diffusion models have set new benchmarks in image and video generation, enabling realistic visual synthesis across single- and multi-frame contexts. However, these models still struggle with efficiently and explicitly generating 3D-consistent content. To address this, we propose World-consistent Video Diffusion (WVD), a novel framework that incorporates explicit 3D supervision using XYZ images, which encode global 3D coordinates for each image pixel. More specifically, we train a diffusion transformer to learn the joint distribution of RGB and XYZ frames. This approach supports multi-task adaptability via a flexible inpainting strategy. For example, WVD can estimate XYZ frames from ground-truth RGB or generate novel RGB frames using XYZ projections along a specified camera trajectory. In doing so, WVD unifies tasks like single-image-to-3D generation, multi-view stereo, and camera-controlled video generation. Our approach demonstrates competitive performance across multiple benchmarks, providing a scalable solution for 3D-consistent video and image generation with a single pretrained model.

****

A Stitch in Time Saves Nine: Small VLM is a Precise Guidance for Accelerating Large VLMs

Wangbo Zhao, Yizeng Han, Jiasheng Tang, Zhikai Li, Yibing Song, Kai Wang, Zhangyang Wang, Yang You; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19814-19824

Vision-language models (VLMs) have shown remarkable success across various multi-modal tasks, yet large VLMs encounter significant efficiency challenges due to processing numerous visual tokens. A promising approach to accelerating large VLM inference is using partial information, such as attention maps from specific layers, to assess token importance and prune less essential tokens. However, our

study reveals three key insights: (i) Partial attention information is insufficient for accurately identifying critical visual tokens, resulting in suboptimal performance, especially at low token retention ratios; (ii) Global attention information, such as the attention map aggregated across all layers, more effectively preserves essential tokens and maintains performance under aggressive pruning. However, it requires a full inference pass, which increases computational load and is therefore impractical in existing methods; and (iii) The global attention map aggregated from a small VLM closely resembles that of a large VLM, suggesting an efficient alternative. Based on these findings, we introduce \underline S mall VLM \underline G uidance for \underline L arge VLMs (SGL). Specifically, we employ the aggregated attention map from a small VLM guide the pruning of visual tokens in a large VLM. Additionally, we develop a small VLM early exiting mechanism to make full use of the small VLM's predictions, dynamically invoking the larger VLM only when necessary, yielding a superior trade-off between accuracy and computational cost. Extensive evaluations across 11 benchmarks demonstrate the effectiveness and generalizability of our method, achieving up to 91% pruning ratio for visual tokens while retaining competitive performance.
****

Zero-Shot Novel View and Depth Synthesis with Multi-View Geometric Diffusion
Vitor Guizilini, Muhammad Zubair Irshad, Dian Chen, Greg Shakhnarovich, Rares Ambrus; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 764-776
Current methods for 3D scene reconstruction from sparse posed images employ intermediate 3D representations such as neural fields, voxel grids, or 3D Gaussians, to achieve multi-view consistent scene appearance and geometry. In this paper we introduce MVGD, a diffusion-based architecture capable of direct pixel-level generation of images and depth maps from novel viewpoints, given an arbitrary number of input views. Our method uses raymap conditioning to both augment visual features with spatial information from different viewpoints, as well as to guide the generation of images and depth maps from novel views. A key aspect of our approach is the multi-task generation of images and depth maps, using learnable task embeddings to guide the diffusion process towards specific modalities. We train this model on a collection of more than 60 million multi-view samples from publicly available datasets, and propose techniques to enable efficient and consistent learning in such diverse conditions. We also propose a novel strategy that enables the efficient training of larger models by incrementally fine-tuning smaller ones, with promising scaling behavior. Through extensive experiments, we report state-of-the-art results in multiple novel view synthesis benchmarks, as well as multi-view stereo and video depth estimation.
****

Unveiling Visual Perception in Language Models: An Attention Head Analysis Approach
Jing Bi, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, Bingjie Wang, Chenliang Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4135-4144
Recent advancements in Multimodal Large Language Models (MLLMs) have demonstrated remarkable progress in visual understanding. This impressive leap raises a compelling question: how can language models, initially trained solely on linguistic data, effectively interpret and process visual content? This paper aims to address this question with systematic investigation across 4 model families and 4 model scales, uncovering a unique class of attention heads that focus specifically on visual content. Our analysis reveals a strong correlation between the behavior of these attention heads, the distribution of attention weights, and their concentration on visual tokens within the input. These findings enhance our understanding of how LLMs adapt to multimodal tasks, demonstrating their potential to bridge the gap between textual and visual understanding. This work paves the way for the development of AI systems capable of engaging with diverse modalities.
****

SemanticDraw: Towards Real-Time Interactive Content Creation from Image Diffusion Models

Jaerin Lee, Daniel Sungho Jung, Kanggeon Lee, Kyoung Mu Lee; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13021-13030
We introduce SemanticDraw, a new paradigm of interactive content creation where high-quality images are generated in near real-time from given multiple hand-drawn regions, each encoding prescribed semantic meaning. In order to maximize the productivity of content creators and to fully realize their artistic imagination, it requires both quick interactive interfaces and fine-grained regional controls in their tools. Despite astonishing generation quality from recent diffusion models, we find that existing approaches for regional controllability are very slow (52 seconds for 512 x 512 image) while not compatible with acceleration methods such as LCM, blocking their huge potential in interactive content creation. From this observation, we build our solution for interactive content creation in two steps: (1) we establish compatibility between region-based controls and acceleration techniques for diffusion models, maintaining high fidelity of multi-prompt image generation with x 10 reduced number of inference steps, (2) we increase the generation throughput with our new multi-prompt stream batch pipeline, enabling low-latency generation from multiple, region-based text prompts on a single RTX 2080 Ti GPU. Our proposed framework is generalizable to any existing diffusion models and acceleration schedulers, allowing sub-second (0.64 seconds) image content creation application upon well-established image diffusion models.
****
SemiDAViL: Semi-supervised Domain Adaptation with Vision-Language Guidance for Semantic Segmentation
Hritam Basak, Zhaozheng Yin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9816-9828
Domain Adaptation (DA) and Semi-supervised Learning (SSL) converge in Semi-supervised Domain Adaptation (SSDA), where the objective is to transfer knowledge from a source domain to a target domain using a combination of limited labeled target samples and abundant unlabeled target data. Although intuitive, a simple amalgamation of DA and SSL is suboptimal in semantic segmentation due to two major reasons: (1) previous methods, while able to learn good segmentation boundaries, are prone to confuse classes with similar visual appearance due to limited supervision; and (2) skewed and imbalanced training data distribution preferring source representation learning whereas impeding from exploring limited information about tailed classes. Language guidance can serve as a pivotal semantic bridge, facilitating robust class discrimination and mitigating visual ambiguities by leveraging the rich semantic relationships encoded in pre-trained language models to enhance feature representations across domains. Therefore, we propose the first language-guided SSDA setting for semantic segmentation in this work. Specifically, we harness the semantic generalization capabilities inherent in vision-language models (VLMs) to establish a synergistic framework within the SSDA paradigm. To address the inherent class-imbalance challenges in long-tailed distributions, we introduce class-balanced segmentation loss formulations that effectively regularize the learning process. Through extensive experimentation across diverse domain adaptation scenarios, our approach demonstrates substantial performance improvements over contemporary state-of-the-art (SoTA) methodologies.
****
Learning Partonomic 3D Reconstruction from Image Collections
Xiaoqian Ruan, Pei Yu, Dian Jia, Hyeonjeong Park, Peixi Xiong, Wei Tang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26734-26744
Reconstructing the 3D shape of an object from a single-view image is a fundamental task in computer vision. Recent advances in differentiable rendering have enabled 3D reconstruction from image collections using only 2D annotations. However, these methods mainly focus on whole-object reconstruction and overlook object partonomy, which is essential for intelligent agents interacting with physical environments. This paper aims at learning partonomic 3D reconstruction from collections of images with only 2D annotations. Our goal is not only to reconstruct the shape of an object from a single-view image but also to decompose the shape into meaningful semantic parts. To handle the expanded solution space and frequen

t part occlusions in single-view images, we introduce a novel approach that repr
esents, parses, and learns the structural compositionality of 3D objects. This a
pproach comprises: (1) a compact and expressive compositional representation of
object geometry, achieved through disentangled modeling of large shape variation
s, constituent parts, and detailed part deformations as multi-granularity neural
 fields; (2) a part transformer that recovers precise partonomic geometry and ha
ndles occlusions, through effective part-to-pixel grounding and part-to-part rel
ational modeling; and (3) a 2D-supervised learning method that jointly learns th
e compositional representation and part transformer, by bridging object shape an
d parts, image synthesis, and differentiable rendering. Extensive experiments on
 ShapeNetPart, PartNet, and CUB-200-2011 demonstrate the effectiveness of our ap
proach on both overall and partonomic reconstruction. Code, models, and data are
 avaliable at https://github.com/XiaoqianRuan1/Partonomic_Reconstruction.
****

Arc2Avatar: Generating Expressive 3D Avatars from a Single Image via ID Guidance
Dimitrios Gerogiannis, Foivos Paraperas Papantoniou, Rolandos Alexandros Potamia
s, Alexandros Lattas, Stefanos Zafeiriou; Proceedings of the Computer Vision and
 Pattern Recognition Conference (CVPR), 2025, pp. 10770-10782
Inspired by the effectiveness of 3D Gaussian Splatting (3DGS) in reconstructing
detailed 3D scenes within multi-view setups and the emergence of large 2D human
foundation models, we introduce Arc2Avatar, the first SDS-based method utilizing
 a human face foundation model as guidance with just a single image as input. To
 achieve that, we extend such a model for diverse-view human head generation by
fine-tuning on synthetic data and modifying its conditioning. Our avatars mainta
in a dense correspondence with a human face mesh template, allowing blendshape-b
ased expression generation. This is achieved through a modified 3DGS approach, c
onnectivity regularizers, and a strategic initialization tailored for our task.
Additionally, we propose an optional efficient SDS-based correction step to refi
ne the blendshape expressions, enhancing realism and diversity. Experiments demo
nstrate that Arc2Avatar achieves state-of-the-art realism and identity preservat
ion, effectively addressing color issues by allowing the use of very low guidanc
e, enabled by our strong identity prior and initialization strategy, without com
promising detail.
****

Seeing Speech and Sound: Distinguishing and Locating Audio Sources in Visual Sce
nes
Hyeonggon Ryu, Seongyu Kim, Joon Son Chung, Arda Senocak; Proceedings of the Com
puter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13540-13549
We present a unified model capable of simultaneously grounding both spoken langu
age and non-speech sounds within a visual scene, addressing key limitations in c
urrent audio-visual grounding models. Existing approaches are typically limited
to handling either speech or non-speech sounds independently, or at best, togeth
er but sequentially without mixing. This limitation prevents them from capturing
 the complexity of real-world audio sources that are often mixed. Our approach i
ntroduces a "mix-and-separate" framework with audio-visual alignment objectives
that jointly learn correspondence and disentanglement using mixed audio. Through
 these objectives, our model learns to produce distinct embeddings for each audi
o type, enabling effective disentanglement and grounding across mixed audio sour
ces.Additionally, we created a new dataset to evaluate simultaneous grounding of
 mixed audio sources, demonstrating that our model outperforms prior methods. Ou
r approach also achieves state-of-the-art performance in standard segmentation a
nd cross-modal retrieval tasks, highlighting the benefits of our mix-and-separat
e approach.
****

Structure from Collision
Takuhiro Kaneko; Proceedings of the Computer Vision and Pattern Recognition Conf
erence (CVPR), 2025, pp. 16314-16324
Recent advancements in neural 3D representations, such as neural radiance fields
 (NeRF) and 3D Gaussian splatting (3DGS), have made accurate estimation of the 3
D structure from multiview images possible. However, this capability is limited

to estimating the visible external structure, and it is still difficult to identify the invisible internal structure hidden behind the surface. To overcome this limitation, we address a new task called structure from collision (SfC), which aims to estimate the structure (including the invisible internal one) of an object from the appearance changes at collision. To solve this task, we propose a novel model called SfC-NeRF, which optimizes the invisible internal structure of the object through a video sequence under physical, appearance (i.e., visible external structure)-preserving, and keyframe constraints. In particular, to avoid falling into undesirable local optima owing to its ill-posed nature, we propose volume annealing, i.e., searching for the global optima by repeatedly reducing and expanding the volume. Extensive experiments on 115 objects involving diverse structures (i.e., various cavity shapes, locations, and sizes) and various material properties reveal the properties of SfC and demonstrate the effectiveness of the proposed SfC-NeRF.

****

ODA-GAN: Orthogonal Decoupling Alignment GAN Assisted by Weakly-supervised Learning for Virtual Immunohistochemistry Staining

Tong Wang, Mingkang Wang, Zhongze Wang, Hongkai Wang, Qi Xu, Fengyu Cong, Hongming Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25920-25929

Recently, virtual staining has emerged as a promising alternative to revolutionize histological staining by digitally generating stains. However, most existing methods suffer from the curse of staining unreality and unreliability. In this paper, we propose the Orthogonal Decoupling Alignment Generative Adversarial Network (ODA-GAN) for unpaired virtual immunohistochemistry (IHC) staining. Our approach is based on the assumption that an image consists of IHC staining-related features, which influence staining distribution and intensity, and staining-unrelated features, such as tissue morphology. Leveraging a pathology foundation model, we first develop a weakly-supervised segmentation pipeline as an alternative to expert annotations. We introduce an Orthogonal MLP (O-MLP) module to project image features into an orthogonal space, decoupling them into staining-related and unrelated components. Additionally, we propose a Dual-stream PatchNCE (DPNCE) loss to resolve contrastive learning contradictions in the staining-related space, thereby enhancing staining accuracy. To further improve realism, we introduce a Multi-layer Domain Alignment (MDA) module to bridge the domain gap between generated and real IHC images. Evaluations on three benchmark datasets show that our ODA-GAN reaches state-of-the-art (SOTA) performance. Our source code is available at https://github.com/ittong/ODA-GAN.

****

EVOS: Efficient Implicit Neural Training via EVOlutionary Selector

Weixiang Zhang, Shuzhao Xie, Chengwei Ren, Siyi Xie, Chen Tang, Shijia Ge, Mingzi Wang, Zhi Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30472-30482

We propose EVOlutionary Selector (EVOS), an efficient training paradigm for accelerating Implicit Neural Representation (INR). Unlike conventional INR training that feeds all samples through the neural network in each iteration, our approach restricts training to strategically selected points, reducing computational overhead by eliminating redundant forward passes.Specifically, we treat each sample as an individual in an evolutionary process, where only those fittest ones survive and merit inclusion in training, adaptively evolving with the neural network dynamics. While this is conceptually similar to Evolutionary Algorithms, their distinct objectives (selection for acceleration vs. iterative solution optimization) require a fundamental redefinition of evolutionary mechanisms for our context.In response, we design sparse fitness evaluation, frequency-guided crossover, and augmented unbiased mutation to comprise EVOS. These components respectively guide sample selection with reduced computational cost, enhance performance through frequency-domain balance, and mitigate selection bias from cached evaluation. Extensive experiments demonstrate that our method achieves approximately 48%-66% reduction in training time while ensuring superior convergence without additional cost, establishing state-of-the-art acceleration among recent sampling-ba

sed strategies.
****
Crab: A Unified Audio-Visual Scene Understanding Model with Explicit Cooperation
Henghui Du, Guangyao Li, Chang Zhou, Chunjie Zhang, Alan Zhao, Di Hu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18804-18814
In recent years, numerous tasks have been proposed to encourage model to develop specified capability in understanding audio-visual scene, primarily categorized into temporal localization, spatial localization, spatio-temporal reasoning, and pixel-level understanding. Instead, human possesses a unified understanding ability for diversified tasks. Therefore, designing an audio-visual model with general capability to unify these tasks is of great value. However, simply joint training for all tasks can lead to interference due to the heterogeneity of audiovisual data and complex relationship among tasks. We argue that this problem can be solved through explicit cooperation among tasks. To achieve this goal, we propose a unified learning method which achieves explicit inter-task cooperation from both the perspectives of data and model thoroughly. Specifically, considering the labels of existing datasets are simple words, we carefully refine these datasets and construct an Audio-Visual Unified Instruction-tuning dataset with Explicit reasoning process (AV-UIE), which clarifies the cooperative relationship among tasks. Subsequently, to facilitate concrete cooperation in learning stage, an interaction-aware LoRA structure with multiple LoRA heads is designed to learn different aspects of audiovisual data interaction. By unifying the explicit cooperation across the data and model aspect, our method not only surpasses existing unified audio-visual model on multiple tasks, but also outperforms most specialized models for certain tasks. Furthermore, we also visualize the process of explicit cooperation and surprisingly find that each LoRA head has certain audio-visual understanding ability. Code and dataset: https://github.com/GeWu-Lab/Crab
****
Nullu: Mitigating Object Hallucinations in Large Vision-Language Models via HalluSpace Projection
Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, Chao Shen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14635-14645
Recent studies have shown that large vision-language models (LVLMs) often suffer from the issue of object hallucinations (OH). To mitigate this issue, we introduce an efficient method that edits the model weights based on an unsafe subspace, which we call HalluSpace in this paper. With truthful and hallucinated text prompts accompanying the visual content as inputs, the HalluSpace can be identified by extracting the hallucinated embedding features and removing the truthful representations in LVLMs. By orthogonalizing the model weights, input features will be projected into the Null space of the HalluSpace to reduce OH, based on which we name our method Nullu. We reveal that HalluSpaces generally contain prior information in the large language models (LLMs) applied to build LVLMs, which have been shown as essential causes of OH in previous studies. Therefore, null space projection suppresses the LLMs' priors to filter out the hallucinated features, resulting in contextually accurate outputs. Experiments show that our method can effectively mitigate OH across different LVLM families without extra inference costs and also show strong performance in general LVLM benchmarks. Code is released at https://github.com/Ziwei-Zheng/Nullu.
****
OralXrays-9: Towards Hospital-Scale Panoramic X-ray Anomaly Detection via Personalized Multi-Object Query-Aware Mining
Bingzhi Chen, Sisi Fu, Xiaocheng Fang, Jieyi Cai, Boya Zhang, Minhua Lu, Yishu Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15570-15579
In clinical practice, panoramic dental radiography is a widely employed imaging technique that can provide a detailed and comprehensive view of dental structures and surrounding tissues for identifying various oral anomalies. However, due to the complexity of oral anomalies and the scarcity of available data, existing

research still suffers from substantial challenges in automated oral anomaly det
ection. To this end, this paper presents a new hospital-scale panoramic X-ray be
nchmark, namely "OralXrays-9", which consists of 12,688 panoramic X-ray images w
ith 84,113 meticulously annotated instances across nine common oral anomalies. C
orrespondingly, we propose a personalized Multi-Object Query-Aware Mining (MOQAM
) paradigm, which jointly incorporates the Distribution-IoU Region Proposal Netw
ork (DI-RPN) and Class-Balanced Spherical Contrastive Regularization (CB-SCR) me
chanisms to address the challenges posed by multi-scale variations and class-imb
alanced distributions. To the best of our knowledge, this is the first attempt t
o develop AI-driven diagnostic systems specifically designed for multi-object or
al anomaly detection, utilizing publicly available data resources. Extensive exp
eriments on the newly-published OralXrays-9 dataset and real-world nature scenar
ios consistently demonstrate the superiority of our MOQAM in revolutionizing ora
l healthcare practices.
****

MEET: Towards Memory-Efficient Temporal Sparse Deep Neural Networks
Zeqi Zhu, Ibrahim Batuhan Akkaya, Luc Waeijen, Egor Bondarev, Arash Pourtaherian
, Orlando Moreira; Proceedings of the Computer Vision and Pattern Recognition Co
nference (CVPR), 2025, pp. 29309-29320
Deep Neural Networks (DNNs) are accurate but compute-intensive, leading to subst
antial energy consumption during inference. Exploiting temporal redundancy throu
gh \Delta-\Sigma convolution in video processing has proven to greatly enhance c
omputation efficiency. However, temporal \Delta-\Sigma DNNs typically require su
bstantial memory for storing neuron states to compute inter-frame differences, h
indering their on-chip deployment. To mitigate this memory cost, directly compre
ssing the states can disrupt the linearity of temporal \Delta-\Sigma convolution
, causing accumulated errors in long-term \Delta-\Sigma processing. Thus, we pro
pose MEET, an optimization framework for MEmory-Efficient Temporal \Delta-\Sigma
 DNNs. MEET transfers the state compression challenge to a well-established weig
ht compression problem by trading fewer activations for more weights and introdu
ces a co-design of network architecture and suppression method to optimize for m
ixed spatial-temporal execution. Evaluations on three vision applications demons
trate a reduction of 5.1~13.3 xin total memory compared to the most computation-
efficient temporal DNNs, while preserving the computation efficiency and model a
ccuracy in long-term \Delta-\Sigma processing. MEET facilitates the deployment o
f temporal \Delta-\Sigma DNNs within on-chip memory of embedded event-driven pla
tforms, empowering low-power edge processing.
****

SplatAD: Real-Time Lidar and Camera Rendering with 3D Gaussian Splatting for Aut
onomous Driving
Georg Hess, Carl Lindström, Maryam Fatemi, Christoffer Petersson, Lennart Svenss
on; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)
, 2025, pp. 11982-11992
Ensuring the safety of autonomous robots, such as self-driving vehicles, require
s extensive testing across diverse driving scenarios. Simulation is a key ingred
ient for conducting such testing in a cost-effective and scalable way. Neural re
ndering methods have gained popularity, as they can build simulation environment
s from collected logs in a data-driven manner. However, existing neural radiance
 field (NeRF) methods for sensor-realistic rendering of camera and lidar data su
ffer from low rendering speeds, limiting their applicability for large-scale tes
ting. While 3D Gaussian Splatting (3DGS) enables real-time rendering, current me
thods are limited to camera data and are unable to render lidar data essential f
or autonomous driving. To address these limitations, we propose SplatAD, the fir
st 3DGS-based method for realistic, real-time rendering of dynamic scenes for bo
th camera and lidar data. SplatAD accurately models key sensor-specific phenomen
a such as rolling shutter effects, lidar intensity, and lidar ray dropouts, usin
g purpose-built algorithms to optimize rendering efficiency. Evaluation across t
hree autonomous driving datasets demonstrates that SplatAD achieves state-of-the
-art rendering quality with up to +2 PSNR for NVS and +3 PSNR for reconstruction
 while increasing rendering speed over NeRF-based methods by an order of magnitu

de. Code to be released upon publication.
****

Audio-Visual Instance Segmentation
Ruohao Guo, Xianghua Ying, Yaru Chen, Dantong Niu, Guangyao Li, Liao Qu, Yanyu Qi, Jinxing Zhou, Bowei Xing, Wenzhen Yue, Ji Shi, Qixun Wang, Peiliang Zhang, Buwen Liang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13550-13560
In this paper, we propose a new multi-modal task, termed audio-visual instance segmentation (AVIS), which aims to simultaneously identify, segment and track individual sounding object instances in audible videos. To facilitate this research, we introduce a high-quality benchmark named AVISeg, containing over 90K instance masks from 26 semantic categories in 926 long videos. Additionally, we propose a strong baseline model for this task. Our model first localizes sound source within each frame, and condenses object-specific contexts into concise tokens. Then it builds long-range audio-visual dependencies between these tokens using window-based attention, and tracks sounding objects among the entire video sequences. Extensive experiments reveal that our method performs best on AVISeg, surpassing the existing methods from related tasks. We further conduct the evaluation on several multi-modal large models. Unfortunately, they exhibits subpar performance on instance-level sound source localization and temporal perception. We expect that AVIS will inspire the community towards a more comprehensive multi-modal understanding. Dataset and code is available at https://github.com/ruohaoguo/avis.
****

Probabilistic Prompt Distribution Learning for Animal Pose Estimation
Jiyong Rao, Brian Nlong Zhao, Yu Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29438-29447
Multi-species animal pose estimation has emerged as a challenging yet critical task, hindered by substantial visual diversity and uncertainty. This paper challenges the problem by efficient prompt learning for Vision-Language Pretrained (VLP) models, e.g. CLIP, aiming to resolve the cross-species generalization problem. At the core of the solution lies in the prompt designing, probabilistic prompt modeling and cross-modal adaptation, thereby enabling prompts to compensate for cross-modal information and effectively overcome large data variances under unbalanced data distribution. To this end, we propose a novel probabilistic prompting approach to fully explore textual descriptions, which could alleviate the diversity issues caused by long-tail property and increase the adaptability of prompts on unseen category instance. Specifically, we first introduce a set of learnable prompts and propose a diversity loss to maintain distinctiveness among prompts, thus representing diverse image attributes. Diverse textual probabilistic representations are sampled and used as the guidance for the pose estimation. Subsequently, we explore three different cross-modal fusion strategies at spatial level to alleviate the adverse impacts of visual uncertainty. Extensive experiments on multi-species animal pose benchmarks show that our method achieves the state-of-the-art performance under both supervised and zero-shot settings.
****

dFLMoE: Decentralized Federated Learning via Mixture of Experts for Medical Data Analysis
Luyuan Xie, Tianyu Luan, Wenyuan Cai, Guochen Yan, Zhaoyu Chen, Nan Xi, Yuejian Fang, Qingni Shen, Zhonghai Wu, Junsong Yuan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10203-10213
Federated learning has wide applications in the medical field. It enables knowledge sharing among different healthcare institutes while protecting patients' privacy. However, existing federated learning systems are typically centralized, requiring clients to upload client-specific knowledge to a central server for aggregation. This centralized approach would integrate the knowledge from each client into a centralized server, and the knowledge would be already undermined during the centralized integration before it reaches back to each client. Besides, the centralized approach also creates a dependency on the central server, which may affect training stability if the server malfunctions or connections are unstab

le. To address these issues, we propose a decentralized federated learning frame work named dFLMoE. In our framework, clients directly exchange lightweight head models with each other. After exchanging, each client treats both local and rece ived head models as individual experts, and utilizes a client-specific Mixture o f Experts (MoE) approach to make collective decisions. This design not only redu ces the knowledge damage with client-specific aggregations but also removes the dependency on the central server to enhance the robustness of the framework. We validate our framework on multiple medical tasks, demonstrating that our method evidently outperforms state-of-the-art approaches under both model homogeneity a nd heterogeneity settings.
****

Reconstructing Humans with a Biomechanically Accurate Skeleton
Yan Xia, Xiaowei Zhou, Etienne Vouga, Qixing Huang, Georgios Pavlakos; Proceedin gs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5 355-5365
In this paper, we introduce a method for reconstructing 3D humans from a single image using a biomechanically accurate skeleton model. To achieve this, we train a transformer that takes an image as input and estimates the parameters of the model. Due to the lack of training data for this task, we build a pipeline to pr oduce pseudo ground truth model parameters for single images and implement a tra ining procedure that iteratively refines these pseudo labels. Compared to state- of-the-art methods for 3D human mesh recovery, our model achieves competitive pe rformance on standard benchmarks, while it significantly outperforms them in set tings with extreme 3D poses and viewpoints. Additionally, we show that previous reconstruction methods frequently violate joint angle limits, leading to unnatur al rotations. In contrast, our approach leverages the biomechanically plausible degrees of freedom making more realistic joint rotation estimates. We validate o ur approach across multiple human pose estimation benchmarks. We make the code, models and data available at: https://isshikihugh.github.io/HSMR/
****

AdaCM^2: On Understanding Extremely Long-Term Video with Adaptive Cross-Modality Memory Reduction
Yuanbin Man, Ying Huang, Chengming Zhang, Bingzhe Li, Wei Niu, Miao Yin; Proceed ings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8534-8544
The advancements in large language models (LLMs) have propelled the improvement of video understanding tasks by incorporating LLMs with visual models. However, most existing LLM-based models (e.g., VideoLLaMA, VideoChat) are constrained to processing short-duration videos. Recent attempts to understand long-term videos by extracting and compressing visual features into a fixed memory size. Neverth eless, those methods leverage only visual modality to merge video tokens and ove rlook the correlation between visual and textual queries, leading to difficultie s in effectively handling complex question-answering tasks. To address the chall enges of long videos and complex prompts, we propose AdaCM^2, which, for the fir st time, introduces an adaptive cross-modality memory reduction approach to vide o-text alignment in an auto-regressive manner on video streams. Our extensive ex periments on various video understanding tasks, such as video captioning, video question answering, and video classification, demonstrate that AdaCM^2 achieves state-of-the-art performance across multiple datasets while significantly reduci ng memory usage. Notably, it achieves a 4.5% improvement across multiple tasks i n the LVU dataset with a GPU memory consumption reduction of up to 65%.
****

Mitigating Object Hallucinations in Large Vision-Language Models with Assembly o f Global and Local Attention
Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, Qianying Wang, Ping C hen, Xiaoqin Zhang, Shijian Lu; Proceedings of the Computer Vision and Pattern R ecognition Conference (CVPR), 2025, pp. 29915-29926
Despite great success across various multimodal tasks, Large Vision-Language Mod els (LVLMs) often encounter object hallucinations with generated textual respons es being inconsistent with the actual objects in images. We examine different LV

LMs and pinpoint that one root cause of object hallucinations lies with deficien t attention on discriminative image features. Specifically, LVLMs often predomin antly attend to prompt-irrelevant global features instead of prompt-relevant loc al features, undermining their visual grounding capacity and leading to object h allucinations. We propose Assembly of Global and Local Attention (AGLA) , a trai ning-free and plug-and-play approach that mitigates hallucinations by assembling global features for response generation and local features for visual discrimin ation simultaneously. Specifically, we introduce an image-prompt matching schem e that captures prompt-relevant local features from images, leading to an augmen ted view of the input image where prompt-relevant content is highlighted while i rrelevant distractions are suppressed. Hallucinations can thus be mitigated wit h a calibrated logit distribution that is from generative global features of the original image and discriminative local features of the augmented image. Extens ive experiments show the superiority of AGLA in LVLM hallucination mitigation, d emonstrating its wide applicability across both discriminative and generative ta sks. Our code is available at https://github.com/Lackel/AGLA.

****

VGGT: Visual Geometry Grounded Transformer

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, David Novotny; Proceedings of the Computer Vision and Pattern Recognition Confe rence (CVPR), 2025, pp. 5294-5306

We present VGGT, a feed-forward neural network that directly infers all key 3D a ttributes of a scene, including camera parameters, point maps, depth maps, and 3 D point tracks, from one, a few, or hundreds of its views. This approach is a st ep forward in 3D computer vision, where models have typically been constrained t o and specialized for single tasks. It is also simple and efficient, reconstruct ing images in under one second, and still outperforming alternatives that requir e post-processing with visual geometry optimization techniques. The network achi eves state-of-the-art results in multiple 3D tasks, including camera parameter e stimation, multi-view depth estimation, dense point cloud reconstruction, and 3D point tracking. We also show that using pretrained VGGT as a feature backbone s ignificantly enhances downstream tasks, such as non-rigid point tracking and fee d-forward novel view synthesis. Code and models are publicly available at https: //github.com/facebookresearch/vggt.

****

Silent Branding Attack: Trigger-free Data Poisoning Attack on Text-to-Image Diff usion Models

Sangwon Jang, June Suk Choi, Jaehyeong Jo, Kimin Lee, Sung Ju Hwang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 820 3-8212

Text-to-image diffusion models have achieved remarkable success in generating hi gh-quality contents from text prompts. However, their reliance on publicly avail able data and the growing trend of data sharing for fine-tuning make these model s particularly vulnerable to data poisoning attacks. In this work, we introduce the Silent Branding Attack, a novel data poisoning method that manipulates text- to-image diffusion models to generate images containing specific brand logos or symbols without any text triggers. We find that when certain visual patterns are repeatedly in the training data, the model learns to reproduce them naturally i n its outputs, even without prompt mentions. Leveraging this, we develop an auto mated data poisoning algorithm that unobtrusively injects logos into original im ages, ensuring they blend naturally and remain undetected. Models trained on thi s poisoned dataset generate images containing logos without degrading image qual ity or text alignment. We experimentally validate our silent branding attack acr oss two realistic settings on large-scale high-quality image datasets and style personalization datasets, achieving high success rates even without a specific t ext trigger. Human evaluation and quantitative metrics including logo detection show that our method can stealthily embed logos.

****

UniSTD: Towards Unified Spatio-Temporal Learning across Diverse Disciplines

Chen Tang, Xinzhu Ma, Encheng Su, Xiufeng Song, Xiaohong Liu, Wei-Hong Li, Lei B

ai, Wanli Ouyang, Xiangyu Yue; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29213-29224
Traditional spatiotemporal models generally rely on task-specific architectures, which limit their generalizability and scalability across diverse tasks due to domain-specific design requirements. In this paper, we introduce UniSTD, a unified Transformer-based framework for spatiotemporal modeling, which is inspired by advances in recent foundation models with the two-stage pretraining-then-adaption paradigm. Specifically, our work demonstrates that task-agnostic pretraining on 2D vision and vision-text datasets can build a generalizable model foundation for spatiotemporal learning, followed by specialized joint training on spatiotemporal datasets to enhance task-specific adaptability. To improve the learning capabilities across domains, our framework employs a rank-adaptive mixture-of-expert adaptation by using fractional interpolation to relax the discrete variables so that can be optimized in the continuous space. Additionally, we introduce a temporal module to incorporate temporal dynamics explicitly. We evaluate our approach on a large-scale dataset covering 10 tasks across 4 disciplines, demonstrating that a unified spatiotemporal model can achieve scalable, cross-task learning and support up to 10 tasks simultaneously within one model while reducing training costs in multi-domain applications. Code will be available at https://github.com/1hunters/UniSTD.
****

Visual Consensus Prompting for Co-Salient Object Detection
Jie Wang, Nana Yu, Zihao Zhang, Yahong Han; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9591-9600
Existing co-salient object detection (CoSOD) methods generally employ a three-stage architecture (i.e., encoding, consensus extraction & dispersion, and prediction) along with a typical full fine-tuning paradigm. Although they yield certain benefits, they exhibit two notable limitations: 1) This architecture relies on encoded features to facilitate consensus extraction, but the meticulously extracted consensus does not provide timely guidance to the encoding stage. 2) This paradigm involves globally updating all parameters of the model, which is parameter-inefficient and hinders the effective representation of knowledge within the foundation model for this task. Therefore, in this paper, we propose an interaction-effective and parameter-efficient concise architecture for the CoSOD task, addressing two key limitations. It introduces, for the first time, a parameter-efficient prompt tuning paradigm and seamlessly embeds consensus into the prompts to formulate task-specific Visual Consensus Prompts (VCP). Our VCP aims to induce the frozen foundation model to perform better on CoSOD tasks by formulating task-specific visual consensus prompts with minimized tunable parameters. Concretely, the primary insight of the purposeful Consensus Prompt Generator (CPG) is to enforce limited tunable parameters to focus on co-salient representations and generate consensus prompts. The formulated Consensus Prompt Disperser (CPD) leverages consensus prompts to form task-specific visual consensus prompts, thereby arousing the powerful potential of pre-trained models in addressing CoSOD tasks. Extensive experiments demonstrate that our concise VCP outperforms 13 cutting-edge full fine-tuning models, achieving the new state of the art (with 6.8% improvement in F_m metrics on the most challenging CoCA dataset). Source code has been available at https://github.com/WJ-CV/VCP.
****

Mani-GS: Gaussian Splatting Manipulation with Triangular Mesh
Xiangjun Gao, Xiaoyu Li, Yiyu Zhuang, Qi Zhang, Wenbo Hu, Chaopeng Zhang, Yao Yao, Ying Shan, Long Quan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21392-21402
Neural 3D representations, such as Neural Radiation Fields (NeRF), excel at producing photorealistic rendering results but lack the flexibility for manipulation and editing which is crucial for content creation. However, manipulating NeRF is not highly controllable and requires a long training and inference time. With the emergence of 3D Gaussian Splatting (3DGS), extremely high-fidelity novel view synthesis can be achieved using an explicit point-based 3D representation with much faster training and rendering speed. However, there is still a lack of eff

ective means to manipulate 3DGS freely while maintaining rendering quality. In t
his work, we aim to tackle the challenge of achieving manipulable photo-realisti
c rendering. We propose to utilize a triangular mesh to manipulate 3DGS directly
 with self-adaptation. This approach reduces the need to design various algorith
ms for different types of 3DGS manipulation. By utilizing a triangle shape-aware
 Gaussian binding and adapting method, we can achieve 3DGS manipulation and pres
erve high-fidelity rendering. In addition, our method is also effective with ina
ccurate meshes extracted from 3DGS. Experiments demonstrate our method's effecti
veness and superiority over baseline approaches.
****

UniHOPE: A Unified Approach for Hand-Only and Hand-Object Pose Estimation
Yinqiao Wang, Hao Xu, Pheng-Ann Heng, Chi-Wing Fu; Proceedings of the Computer V
ision and Pattern Recognition Conference (CVPR), 2025, pp. 12231-12241
Estimating the 3D pose of hand and potential hand-held object from monocular ima
ges is a longstanding challenge. Yet, existing methods are specialized, focusing
 on either bare-hand or hand interacting with object. No method can flexibly han
dle both scenarios and their performance degrades when applied to the other scen
ario. In this paper, we propose UniHOPE, a unified approach for general 3D hand-
object pose estimation, flexibly adapting both scenarios. Technically, we design
 a grasp-aware feature fusion module to integrate hand-object features with an o
bject switcher to dynamically control the hand-object pose estimation according
to grasping status. Further, to uplift the robustness of hand pose estimation re
gardless of object presence, we generate realistic de-occluded image pairs to tr
ain the model to learn object-induced hand occlusions, and formulate multi-level
 feature enhancement techniques for learning occlusion-invariant features. Exten
sive experiments on three commonly-used benchmarks demonstrate UniHOPE's SOTA pe
rformance in addressing hand-only and hand-object scenarios. Code will be releas
ed on https://github.com/JoyboyWang/UniHOPE_Pytorch.
****

RL-RC-DoT: A Block-level RL agent for Task-Aware Video Compression
Uri Gadot, Assaf Shocher, Shie Mannor, Gal Chechik, Assaf Hallak; Proceedings of
 the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12533-
12542
Video encoders optimize compression for human perception by minimizing reconstru
ction error under bit-rate constraints. In many modern applications such as auto
nomous driving, an overwhelming majority of videos serve as input for AI systems
 performing tasks like object recognition or segmentation, rather than being wat
ched by humans. It is therefore useful to optimize the encoder for a downstream
task instead of for perceptual image quality. However, a major challenge is how
to combine such downstream optimization with existing standard video encoders, w
hich are highly efficient and popular. Here, we address this challenge by contro
lling the Quantization Parameters (QPs) at the macro-block level to optimize the
 downstream task. This granular control allows us to prioritize encoding for tas
k-relevant regions within each frame. We formulate this optimization problem as
a Reinforcement Learning (RL) task, where the agent learns to balance long-term
implications of choosing QPs on both task performance and bit-rate constraints.
Notably, our policy does not require the downstream task as an input during infe
rence, making it suitable for streaming applications and edge devices such as ve
hicles. We demonstrate significant improvements in two tasks, car detection, and
 ROI (saliency) encoding. Our approach improves task performance for a given bit
 rate compared to traditional task agnostic encoding methods, paving the way for
 more efficient task-aware video compression.
****

Quantization without Tears
Minghao Fu, Hao Yu, Jie Shao, Junjie Zhou, Ke Zhu, Jianxin Wu; Proceedings of th
e Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4462-4472
Deep neural networks, while achieving remarkable success across diverse tasks, d
emand significant resources, including computation, GPU memory, bandwidth, stora
ge, and energy. Network quantization, as a standard compression and acceleration
 technique, reduces storage costs and enables potential inference acceleration b

y discretizing network weights and activations into a finite set of integer valu
es. However, current quantization methods are often complex and sensitive, requi
ring extensive task-specific hyperparameters, where even a single misconfigurati
on can impair model performance, limiting generality across different models and
 tasks. In this paper, we propose Quantization without Tears (QwT), a method tha
t simultaneously achieves quantization speed, accuracy, simplicity, and generali
ty. The key insight of QwT is to incorporate a lightweight additional structure
into the quantized network to mitigate information loss during quantization. Thi
s structure consists solely of a small set of linear layers, keeping the method
simple and efficient. More importantly, it provides a closed-form solution, allo
wing us to improve accuracy effortlessly under 2 minutes. Extensive experiments
across various vision, language, and multimodal tasks demonstrate that QwT is bo
th highly effective and versatile. In fact, our approach offers a robust solutio
n for network quantization that combines simplicity, accuracy, and adaptability,
 which provides new insights for the design of novel quantization paradigms.
****

PHGC: Procedural Heterogeneous Graph Completion for Natural Language Task Verifi
cation in Egocentric Videos

Xun Jiang, Zhiyi Huang, Xing Xu, Jingkuan Song, Fumin Shen, Heng Tao Shen; Proce
edings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, p
p. 8615-8624

Natural Language-based Egocentric Task Verification (NLETV) aims to equip agents
 to determine if operation flows of procedural tasks in egocentric videos align
with natural language instructions. Describing rules with natural language provi
des generalizable applications, but also raises cross-modal heterogeneity and hi
erarchical misalignment challenges. In this paper, we proposed a novel approach
termed Procedural Heterogeneous Graph Completion (PHGC), which addresses these c
hallenges with heterogeneous graphs representing the logic in rules and operatio
n flows. Specifically, our PHGC method mainly consists of three key components:
(1) Heterogeneous Graph Construction module that defines objective states and op
eration flows as vertices, with temporal and sequential relations as edges. (2)
Cross-Modal Path Finding module that aligns semantic relations between hierarchi
cal video and text elements. (3) Discriminative Entity Representation module exc
avates hidden entities that integrate general logical relations and discriminati
ve cues to reveal final verification results. Additionally, we further construct
ed a new dataset called CSV-NL comprised of realistic videos. Extensive experime
nts on the two benchmark datasets covering both digital and physical scenarios,
i.e., EgoTV and CSV-NL, demonstrate that our proposed PHGC establishes state-of-
the-art performance across different settings. Our code and dataset are availabl
e at https://github.com/XunCHN/PHGC.
****

Recognition-Synergistic Scene Text Editing

Zhengyao Fang, Pengyuan Lyu, Jingjing Wu, Chengquan Zhang, Jun Yu, Guangming Lu,
 Wenjie Pei; Proceedings of the Computer Vision and Pattern Recognition Conferen
ce (CVPR), 2025, pp. 13104-13113

Scene text editing aims to modify text content within scene images while maintai
ning style consistency. Traditional methods achieve this by explicitly disentang
ling style and content from the source image and then fusing the style with the
target content, while ensuring content consistency using a pre-trained recogniti
on model. Despite notable progress, these methods suffer from complex pipelines,
 leading to suboptimal performance in complex scenarios. In this work, we introd
uce Recognition-Synergistic Scene Text Editing (RS-STE), a novel approach that f
ully exploits the intrinsic synergy of text recognition for editing. Our model s
eamlessly integrates text recognition with text editing within a unified framewo
rk, and leverages the recognition model's ability to implicitly disentangle styl
e and content while ensuring content consistency. Specifically, our approach emp
loys a multi-modal parallel decoder based on transformer architecture, which pre
dicts both text content and stylized images in parallel. Additionally, our cycli
c self-supervised fine-tuning strategy enables effective training on unpaired re
al-world data without ground truth, enhancing style and content consistency thro

ugh a twice-cyclic generation process. Built on a relatively simple architecture, RS-STE achieves state-of-the-art performance on both synthetic and real-world benchmarks, and further demonstrates the effectiveness of leveraging the generated hard cases to boost the performance of downstream recognition tasks. Code will be made publicly available.

****

Towards Consistent Multi-Task Learning: Unlocking the Potential of Task-Specific Parameters

Xiaohan Qin, Xiaoxing Wang, Junchi Yan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10067-10076

Multi-task learning (MTL) has gained widespread application for its ability to transfer knowledge across tasks, improving resource efficiency and generalization. However, gradient conflicts from different tasks remain a major challenge in MTL. Previous gradient-based and loss-based methods primarily focus on gradient optimization in shared parameters, often overlooking the potential of task-specific parameters. This work points out that task-specific parameters not only capture task-specific information but also influence the gradients propagated to shared parameters, which in turn affects gradient conflicts. Motivated by this insight, we propose ConsMTL, which models MTL as a bi-level optimization problem: in the upper-level optimization, we perform gradient aggregation on shared parameters to find a joint update vector that minimizes gradient conflicts; in the lower-level optimization, we introduce an additional loss for task-specific parameters guiding the k gradients of shared parameters to gradually converge towards the joint update vector. Our design enables the optimization of both shared and task-specific parameters to consistently alleviate gradient conflicts. Extensive experiments show that ConsMTL achieves state-of-the-art performance across various benchmarks with task numbers ranging from 2 to 40, demonstrating its superior performance.

****

WildAvatar: Learning In-the-wild 3D Avatars from the Web

Zihao Huang, Shoukang Hu, Guangcong Wang, Tianqi Liu, Yuhang Zang, Zhiguo Cao, Wei Li, Ziwei Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15963-15975

Existing research on avatar creation is typically limited to laboratory datasets, which require high costs against scalability and exhibit insufficient representation of the real world. On the other hand, the web abounds with off-the-shelf real-world human videos, but these videos vary in quality and require accurate annotations for avatar creation. To this end, we propose an automatic annotating pipeline with filtering protocols to curate these humans from the web. Our pipeline surpasses state-of-the-art methods on the EMDB benchmark, and the filtering protocols boost verification metrics on web videos. We then curate WildAvatar, a web-scale in-the-wild human avatar creation dataset extracted from YouTube, with 10,000+ different human subjects and scenes. WildAvatar is at least 10xricher than previous datasets for 3D human avatar creation and closer to the real world. To explore its potential, we demonstrate the quality and generalizability of avatar creation methods on WildAvatar. We will publicly release our code, data source links and annotations to push forward 3D human avatar creation and other related fields for real-world applications.

****

BooW-VTON: Boosting In-the-Wild Virtual Try-On via Mask-Free Pseudo Data Training

Xuanpu Zhang, Dan Song, Pengxin Zhan, Tianyu Chang, Jianhao Zeng, Qingguo Chen, Weihua Luo, An-An Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26399-26408

Image-based virtual try-on is an increasingly popular and important task to generate realistic try-on images of the specific person.Recent methods model virtual try-on as image mask-inpaint task, which requires masking the person image and results in significant loss of spatial information. Especially, for in-the-wild try-on scenarios with complex poses and occlusions, mask-based methods often introduce noticeable artifacts. Our research found that a mask-free approach can fu

lly leverage spatial and lighting information from the original person image, en
abling high-quality virtual try-on. Consequently, we propose a novel training pa
radigm for a mask-free try-on diffusion model. We ensure the model's mask-free t
ry-on capability by creating high-quality pseudo-data and further enhance its ha
ndling of complex spatial information through effective in-the-wild data augment
ation. Besides, a try-on localization loss is designed to concentrate on try-on
area while suppressing garment features in non-try-on areas, ensuring precise re
ndering of garments and preservation of fore/back-ground. In the end, we introdu
ce BooW-VTON, the mask-free virtual try-on diffusion model, which delivers SOTA
try-on quality without parsing cost. Extensive qualitative and quantitative expe
riments have demonstrated superior performance in wild scenarios with such a low
-demand input.
****

Rectified Diffusion Guidance for Conditional Generation
Mengfei Xia, Nan Xue, Yujun Shen, Ran Yi, Tieliang Gong, Yong-Jin Liu; Proceedin
gs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1
3371-13380
Classifier-Free Guidance (CFG), which combines the conditional and unconditional
score functions with two coefficients summing to one, serves as a practical tec
hnique for diffusion model sampling. Theoretically, however, denoising with CFG
cannot be expressed as a reciprocal diffusion process, which may consequently le
ave some hidden risks during use. In this work, we revisit the theory behind CFG
and rigorously confirm that the improper configuration of the combination coeff
icients (*i.e.*, the widely used summing-to-one version) brings about expectatio
n shift of the generative distribution. To rectify this issue, we propose ReCFG
with a relaxation on the guidance coefficients such that denoising with ReCFG st
rictly aligns with the diffusion theory. We further show that our approach enjoy
s a **closed-form** solution given the guidance strength. That way, the rectifie
d coefficients can be readily pre-computed via traversing the observed data, lea
ving the sampling speed barely affected. Empirical evidence on real-world data d
emonstrate the compatibility of our post-hoc design with existing state-of-the-a
rt diffusion models, including both class-conditioned ones (*e.g.*, EDM2 on Imag
eNet) and text-conditioned ones (*e.g.*, SD3 on CC12M), without any retraining.
Code is available at https://github.com/thuxmf/recfg.
****

SceneFactor: Factored Latent 3D Diffusion for Controllable 3D Scene Generation
Aleksey Bokhovkin, Quan Meng, Shubham Tulsiani, Angela Dai; Proceedings of the C
omputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 628-639
We present SceneFactor, a diffusion-based approach for large-scale 3D scene gene
ration that enables controllable generation and effortless editing. SceneFactor
enables text-guided 3D scene synthesis through our factored diffusion formulatio
n, leveraging latent semantic and geometric manifolds for generation of arbitrar
y-sized 3D scenes. While text input enables easy, controllable generation, text
guidance remains imprecise for intuitive, localized editing and manipulation of
the generated 3D scenes. Our factored semantic diffusion generates a proxy seman
tic space composed of semantic 3D boxes that enables controllable editing of gen
erated scenes by adding, removing, changing the size of the semantic 3D proxy bo
xes that guides high-fidelity, consistent 3D geometric editing. Extensive experi
ments demonstrate that our approach enables high-fidelity 3D scene synthesis wit
h effective controllable editing through our factored diffusion approach.
****

HiFi-Portrait: Zero-shot Identity-preserved Portrait Generation with High-fideli
ty Multi-face Fusion
Yifang Xu, Benxiang Zhai, Yunzhuo Sun, Ming Li, Yang Li, Sidan Du; Proceedings o
f the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5625-
5635
Recent advancements in diffusion-based technologies have made significant stride
s, particularly in identity-preserved portrait generation (IPG). However, when u
sing multiple reference images from the same ID, existing methods typically prod
uce lower-fidelity portraits and struggle to customize face attributes precisely

. To address these issues, this paper presents HiFi-Portrait, a high-fidelity method for zero-shot portrait generation. Specifically, we first introduce the face refiner and landmark generator to obtain fine-grained multi-face features and 3D-aware face landmarks. The landmarks include the reference ID and the target attributes. Then, we design HiFi-Net to fuse multi-face features and align them with landmarks, which improves ID fidelity and face control. In addition, we devise an automated pipeline to construct an ID-based dataset for training HiFi-Portrait. Extensive experimental results demonstrate that our method surpasses the SOTA approaches in face similarity and controllability. Furthermore, our method is also compatible with previous SDXL-based works.

****

IAAO: Interactive Affordance Learning for Articulated Objects in 3D Environments

Can Zhang, Gim Hee Lee; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12132-12142

This work presents IAAO, a novel framework that builds an explicit 3D model for intelligent agents to gain understanding of articulated objects in their environment through interaction. Unlike prior methods that rely on task-specific networks and assumptions about movable parts, our IAAO leverages large foundation models to estimate interactive affordances and part articulations in three stages. We first build hierarchical features and label fields for each object state using 3D Gaussian Splatting (3DGS) by distilling mask features and view-consistent labels from multi-view images. We then perform object- and part-level queries on the 3D Gaussian primitives to identify static and articulated elements, estimating global transformations and local articulation parameters along with affordances. Finally, scenes from different states are merged and refined based on the estimated transformations, enabling robust affordance-based interaction and manipulation of objects. Experimental results demonstrate the effectiveness of our method.

****

FloVD: Optical Flow Meets Video Diffusion Model for Enhanced Camera-Controlled Video Synthesis

Wonjoon Jin, Qi Dai, Chong Luo, Seung-Hwan Baek, Sunghyun Cho; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2040-2049

We present FloVD, a novel video diffusion model for camera-controllable video generation. FloVD leverages optical flow to represent the motions of the camera and moving objects. This approach offers two key benefits. Since optical flow can be directly estimated from videos, our approach allows for the use of arbitrary training videos without ground-truth camera parameters. Moreover, as background optical flow encodes 3D correlation across different viewpoints, our method enables detailed camera control by leveraging the background motion. To synthesize natural object motion while supporting detailed camera control, our framework adopts a two-stage video synthesis pipeline consisting of optical flow generation and flow-conditioned video synthesis. Extensive experiments demonstrate the superiority of our method over previous approaches in terms of accurate camera control and natural object motion synthesis.

****

RAD: Region-Aware Diffusion Models for Image Inpainting

Sora Kim, Sungho Suh, Minsik Lee; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2439-2448

Diffusion models have achieved remarkable success in image generation, with applications broadening across various domains.Inpainting is one such application that can benefit significantly from diffusion models. Existing methods either hijack the reverse process of a pretrained diffusion model or cast the problem into a larger framework, i.e., conditioned generation. However, these approaches often require nested loops in the generation process or additional components for conditioning. In this paper, we present region-aware diffusion models (RAD) for inpainting with a simple yet effective reformulation of the vanilla diffusion models. RAD utilizes a different noise schedule for each pixel, which allows local regions to be generated asynchronously while considering the global image context. A plain reverse process requires no additional components, enabling RAD to ach

ieve inference time up to 100 times faster than the state-of-the-art approaches. Moreover, we employ low-rank adaptation (LoRA) to fine-tune RAD based on other pretrained diffusion models, reducing computational burdens in training as well. Experiments demonstrated that RAD provides state-of-the-art results both qualitatively and quantitatively, on the FFHQ, LSUN Bedroom, and ImageNet datasets.
****

RaSS: Improving Denoising Diffusion Samplers with Reinforced Active Sampling Scheduler

Xin Ding, Lei Yu, Xin Li, Zhijun Tu, Hanting Chen, Jie Hu, Zhibo Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12923-12933

Recent years have witnessed the great success of denoising diffusion samplers in improving the generative capability and sampling efficiency given a pre-trained diffusion model. However, most sampling schedulers in diffusion models lack the sampling dynamics and planning capability for future generation results, leading to suboptimal solutions. To overcome this, we propose the Reinforced Active Sampling Scheduler, termed RaSS, intending to find the optimal sampling trajectory by actively planning and adjusting the sampling steps for each sampling process in time. Concretely, RaSS divides the whole sampling process into five stages and introduces a reinforcement learning (RL) agent to continuously monitor the generated instance and perceive the potential generation results, thereby achieving optimal instance- and state-adaptive sampling steps decision. Meanwhile, a sampling reward is designed to assist the planning capability of the RL agent by balancing the sampling efficiency and generation quality. The RaSS is a plug-and-play module, which is applicable to multiple denoising diffusion samplers of diffusion models. Extensive experiments on different benchmarks have shown that our RaSS can consistently improve the generation quality and efficiency across various tasks, without introducing significant computational overhead.
****

Supervising Sound Localization by In-the-wild Egomotion

Anna Min, Ziyang Chen, Hang Zhao, Andrew Owens; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23936-23946

We present a method for learning binaural sound localization from ego-motion in videos. When the camera moves in a video, the direction of sound sources will change along with it. We train an audio model to predict sound directions that are consistent with visual estimates of camera motion, which we obtain using methods from multi-view geometry. This provides a weak but plentiful form of supervision that we combine with traditional binaural cues. To evaluate this idea, we propose a dataset of real-world audio-visual videos with ego-motion. We show that our model can successfully learn from this real-world data, and that it obtains strong performance on sound localization tasks.
****

AutoLUT: LUT-Based Image Super-Resolution with Automatic Sampling and Adaptive Residual Learning

Yuheng Xu, Shijie Yang, Xin Liu, Jie Liu, Jie Tang, Gangshan Wu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23131-23140

In recent years, the increasing popularity of Hi-DPI screens has driven a rising demand for high-resolution images. However, the limited computational power of edge devices poses a challenge in deploying complex super-resolution neural networks, highlighting the need for efficient methods. While prior works have made significant progress, they have not fully exploited pixel-level information. Moreover, their reliance on fixed sampling patterns limits both accuracy and the ability to capture fine details in low-resolution images. To address these challenges, we introduce two plug-and-play modules designed to capture and leverage pixel information effectively in Look-Up Table (LUT) based super-resolution networks. Our method introduces Automatic Sampling (AutoSample), a flexible LUT sampling approach where sampling weights are dynamically learned during training to adapt to pixel variations and expand the receptive field without added inference cost. We also incorporate Adaptive Residual Learning (AdaRL) to enhance inter-layer

connections, enabling detailed information flow and improving the network's abi
lity to reconstruct fine details. Our method achieves significant performance im
provements on both MuLUT and SPF-LUT while maintaining similar storage sizes. Sp
ecifically, for MuLUT, we achieve a PSNR improvement of approximately +0.20 dB i
mprovement on average across five datasets . For SPF-LUT, with more than a 50% r
eduction in storage space and about a 2/3 reduction in inference time, our metho
d still maintains performance comparable to the original.
****
Understanding Fine-tuning CLIP for Open-vocabulary Semantic Segmentation in Hype
rbolic Space
Zelin Peng, Zhengqin Xu, Zhilin Zeng, Changsong Wen, Yu Huang, Menglin Yang, Fei
long Tang, Wei Shen; Proceedings of the Computer Vision and Pattern Recognition
Conference (CVPR), 2025, pp. 4562-4572
CLIP, a foundational vision-language model, has emerged as a powerful tool for o
pen-vocabulary semantic segmentation. While freezing the text encoder preserves
its powerful embeddings, recent studies show that fine-tuning both the text and
image encoders jointly significantly enhances segmentation performance, especial
ly for classes from open sets. In this work, we explain this phenomenon from the
perspective of hierarchical alignment, since during fine-tuning, the hierarchy
level of image embeddings shifts from image-level to pixel-level. We achieve thi
s by leveraging hyperbolic space, which naturally encoders hierarchical structur
es. Our key observation is that, during fine-tuning, the hyperbolic radius of CL
IP's text embeddings decreases, facilitating better alignment with the pixel-lev
el hierarchical structure of visual data. Building on this insight, we propose H
yperCLIP, a novel fine-tuning strategy that adjusts the hyperbolic radius of the
text embeddings through scaling transformations. By doing so, HyperCLIP equips
CLIP with segmentation capability while introducing only a small number of learn
able parameters. Our experiments demonstrate that HyperCLIP achieves state-of-th
e-art performance on open-vocabulary semantic segmentation tasks across three be
nchmarks, while fine-tuning only approximately 4% of the total parameters of CLI
P. More importantly, we observe that after adjustment, CLIP's text embeddings ex
hibit a relatively fixed hyperbolic radius across datasets, suggesting that the
segmentation task has a characteristic level in hyperbolic space.
****
TexGaussian: Generating High-quality PBR Material via Octree-based 3D Gaussian S
platting
Bojun Xiong, Jialun Liu, Jiakui Hu, Chenming Wu, Jinbo Wu, Xing Liu, Chen Zhao,
Errui Ding, Zhouhui Lian; Proceedings of the Computer Vision and Pattern Recogni
tion Conference (CVPR), 2025, pp. 551-561
Physically Based Rendering (PBR) materials play a crucial role in modern graphic
s, enabling photorealistic rendering across diverse environment maps. Developing
an effective and efficient algorithm that is capable of automatically generatin
g high-quality PBR materials rather than RGB texture for 3D meshes can significa
ntly streamline the 3D content creation. Most existing methods leverage pre-trai
ned 2D diffusion models for multi-view image synthesis, which often leads to sev
ere inconsistency between the generated textures and input 3D meshes. This paper
presents TexGaussian, a novel method that uses octant-aligned 3D Gaussian Splat
ting for rapid PBR material generation. Specifically, we place each 3D Gaussian
on the finest leaf node of the octree built from the input 3D mesh to render the
multi-view images not only for the albedo map but also for roughness and metall
ic. Moreover, our model is trained in a regression manner instead of diffusion d
enoising, capable of generating the PBR material for a 3D mesh in a single feed-
forward process. Extensive experiments on publicly available benchmarks demonstr
ate that our method synthesizes more visually pleasing PBR materials and runs fa
ster than previous methods in both unconditional and text-conditional scenarios,
exhibiting better consistency with the given geometry. Our code and trained mod
els are available at https://3d-aigc.github.io/TexGaussian.
****
OSV: One Step is Enough for High-Quality Image to Video Generation
Xiaofeng Mao, Zhengkai Jiang, Fu-yun Wang, Jiangning Zhang, Hao Chen, Mingmin Ch

i, Yabiao Wang, Wenhan Luo; Proceedings of the Computer Vision and Pattern Recog nition Conference (CVPR), 2025, pp. 12585-12594

Video diffusion models have shown great potential in generating high-quality vid eos, making them an increasingly popular focus. However, their inherent iterativ e nature leads to substantial computational and time costs. Although techniques such as consistency distillation and adversarial training have been employed to accelerate video diffusion by reducing inference steps, these methods often simp ly transfer the generation approaches from Image diffusion models to video diffu sion models. As a result, these methods frequently fall short in terms of both p erformance and training stability. In this work, we introduce a two-stage traini ng framework that effectively combines consistency distillation with adversarial training to address these challenges. Additionally, we propose a novel video di scriminator design, which eliminates the need for decoding the video latents and improves the final performance. Our model is capable of producing high-quality videos in merely one-step, with the flexibility to perform multi-step refinement for further performance enhancement. Our quantitative evaluation on the OpenVid -1M benchmark shows that our model significantly outperforms existing methods. N otably, our 1-step performance (FVD 171.15) exceeds the 8-step performance of th e consistency distillation based method, AnimateLCM (FVD 184.79), and approache s the 25-step performance of advanced Stable Video Diffusion (FVD 156.94).

****

Fuzzy Multimodal Learning for Trusted Cross-modal Retrieval

Siyuan Duan, Yuan Sun, Dezhong Peng, Zheng Liu, Xiaomin Song, Peng Hu; Proceedin gs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2 0747-20756

Cross-modal retrieval aims to match related samples across distinct modalities, facilitating the retrieval and discovery of heterogeneous information. Although existing methods show promising performance, most are deterministic models and a re unable to capture the uncertainty inherent in the retrieval outputs, leading to potentially unreliable results. To address this issue, we propose a novel fra mework called FUzzy Multimodal lEarning (FUME), which is able to self-estimate e pistemic uncertainty, thereby embracing trusted cross-modal retrieval. Specifica lly, our FUME leverages the Fuzzy Set Theory to view the outputs of the classifi cation network as a set of membership degrees and quantify category credibility by incorporating both possibility and necessity measures. However, directly opti mizing the category credibility could mislead the model by over-optimizing the n ecessity for unmatched categories. To overcome this challenge, we present a nove l fuzzy multimodal learning strategy, which utilizes label information to guide necessity optimization in the right direction, thereby indirectly optimizing cat egory credibility and achieving accurate decision uncertainty quantification. Fu rthermore, we design an uncertainty merging scheme that accounts for decision un certainties, thus further refining uncertainty estimates and boosting the trustw orthiness of retrieval results. Extensive experiments on five benchmark datasets demonstrate that FUME remarkably improves both retrieval performance and reliab ility, offering a prospective solution for cross-modal retrieval in high-stakes applications. Code is available at https://github.com/siyuancncd/FUME .

****

AniGS: Animatable Gaussian Avatar from a Single Image with Inconsistent Gaussian Reconstruction

Lingteng Qiu, Shenhao Zhu, Qi Zuo, Xiaodong Gu, Yuan Dong, Junfei Zhang, Chao Xu , Zhe Li, Weihao Yuan, Liefeng Bo, Guanying Chen, Zilong Dong; Proceedings of th e Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21148-211 58

Generating animatable human avatars from a single image is essential for various digital human modeling applications. Existing 3D reconstruction methods often s truggle to capture fine details in animatable models, while generative approache s for controllable animation, though avoiding explicit 3D modeling, suffer from viewpoint inconsistencies in extreme poses and computational inefficiencies. In this paper, we address these challenges by leveraging the power of generative mo dels to produce detailed multi-view canonical pose images, which help resolve am

biguities in animatable human reconstruction. We then propose a robust method fo r 3D reconstruction of inconsistent images, enabling real-time rendering during inference. Specifically, we adapt a transformer-based Text-to-Video model to gen erate multi-view canonical pose images and normal maps, pretraining on a large-s cale monocular video dataset to improve generalization. To handle view inconsist encies, we recast the reconstruction problem as a 4D task and introduce an effic ient 3D modeling approach using 4D Gaussian Splatting. Experiments demonstrate t hat our method achieves photorealistic, real-time animation of 3D human avatars from in-the-wild images, showcasing its effectiveness and generalization capabil ity.
****

GUI-Xplore: Empowering Generalizable GUI Agents with One Exploration
Yuchen Sun, Shanhui Zhao, Tao Yu, Hao Wen, Samith Va, Mengwei Xu, Yuanchun Li, C hongyang Zhang; Proceedings of the Computer Vision and Pattern Recognition Confe rence (CVPR), 2025, pp. 19477-19486
GUI agents hold significant potential to enhance the experience and efficiency o f human-device interaction. However, current methods face challenges in generali zing across applications (apps) and tasks, primarily due to two fundamental limi tations in existing datasets. First, these datasets overlook developer-induced s tructural variations among apps, limiting the transferability of knowledge acros s diverse software environments. Second, many of them focus solely on navigation tasks, which restricts their capacity to represent comprehensive software archi tectures and complex user interactions. To address these challenges, we introdu ce GUI-Xplore, a dataset meticulously designed to enhance cross-application and cross-task generalization via an exploration-and-reasoning framework. GUI-Xplore integrates pre-recorded exploration videos providing contextual insights, along side five hierarchically structured downstream tasks designed to comprehensively evaluate GUI agent capabilities. To fully exploit GUI-Xplore's unique features, we propose Xplore-Agent, a GUI agent framework that combines Action-aware GUI M odeling with Graph-Guided Environment Reasoning. Further experiments indicate th at Xplore-Agent achieves a 10% improvement over existing methods in unfamiliar e nvironments, yet there remains significant potential for further enhancement tow ards truly generalizable GUI agents.
****

Few-Shot Recognition via Stage-Wise Retrieval-Augmented Finetuning
Tian Liu, Huixin Zhang, Shubham Parashar, Shu Kong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15086-15097
Few-shot recognition (FSR) aims to train a classification model with only a few labeled examples of each concept concerned by a downstream task, where data anno tation cost can be prohibitively high. We develop methods to solve FSR by levera ging a pretrained Vision-Language Model (VLM). We particularly explore retrieval -augmented learning (RAL), which retrieves open data, e.g., the VLM's pretrainin g dataset, to learn models for better serving downstream tasks. RAL has been stu died in zero-shot recognition but remains under-explored in FSR. Although applyi ng RAL to FSR may seem straightforward, we observe interesting and novel challen ges and opportunities. First, somewhat surprisingly, finetuning a VLM on a large amount of retrieved data underperforms state-of-the-art zero-shot methods. This is due to the imbalanced distribution of retrieved data and its domain gaps wit h the few-shot examples in the downstream task. Second, more surprisingly, we fi nd that simply finetuning a VLM solely on few-shot examples significantly outper forms previous FSR methods, and finetuning on the mix of retrieved and few-shot data yields even better results. Third, to mitigate the imbalanced distribution and domain gap issues, we propose Stage-Wise retrieval-Augmented fineTuning (SWA T), which involves end-to-end finetuning on mixed data in the first stage and re training the classifier on the few-shot data in the second stage. Extensive expe riments on nine popular benchmarks demonstrate that SWAT significantly outperfor ms previous methods by >6% accuracy.
****

Concept Replacer: Replacing Sensitive Concepts in Diffusion Models via Precision Localization

Lingyun Zhang, Yu Xie, Yanwei Fu, Ping Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8172-8181
As large-scale diffusion models continue to advance, they excel at producing high-quality images but often generate unwanted content, such as sexually explicit or violent content. Existing methods for concept removal generally guide the image generation process but can unintentionally modify unrelated regions, leading to inconsistencies with the original model. We propose a novel approach for targeted concept replacing in diffusion models, enabling specific concepts to be removed without affecting non-target areas. Our method introduces a dedicated concept localizer for precisely identifying the target concept during the denoising process, trained with few-shot learning to require minimal labeled data. Within the identified region, we introduce a training-free Dual Prompts Cross-Attention (DPCA) module to substitute the target concept, ensuring minimal disruption to surrounding content. We evaluate our method on concept localization precision and replacement efficiency. Experimental results demonstrate that our method achieves superior precision in localizing target concepts and performs coherent concept replacement with minimal impact on non-target areas, outperforming existing approaches.
****

A Regularization-Guided Equivariant Approach for Image Restoration
Yulu Bai, Jiahong Fu, Qi Xie, Deyu Meng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2300-2310
Equivariant and invariant deep learning models have been developed to exploit intrinsic symmetries in data, demonstrating significant effectiveness in certain scenarios. However, these methods often suffer from limited representation accuracy and rely on strict symmetry assumptions that may not hold in practice. These limitations pose a significant drawback for image restoration tasks, which demands high accuracy and precise symmetry representation. To address these challenges, we propose a rotation-equivariant regularization strategy that adaptively enforces the appropriate symmetry constraints on the data while preserving the network's representational accuracy. Specifically, we introduce EQ-Reg, a regularizer designed to enhance rotation equivariance, which innovatively extends the insights of data-augmentation-based and equivariant-based methodologies. This is achieved through self-supervised learning and the spatial rotation and cyclic channel shift of feature maps deduce in the equivariant framework. Our approach firstly enables a non-strictly equivariant network suitable for image restoration, providing a simple and adaptive mechanism for adjusting equivariance based on task. Extensive experiments across three low-level tasks demonstrate the superior accuracy and generalization capability of our method, outperforming state-of-the-art approaches.
****

RestorGS: Depth-aware Gaussian Splatting for Efficient 3D Scene Restoration
Yuanjian Qiao, Mingwen Shao, Lingzhuang Meng, Kai Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11177-11186
3D Gaussian Splatting (3DGS) has recently achieved remarkable progress in novel view synthesis. However, existing methods rely heavily on high-quality data for rendering and struggle to handle degraded scenes with multi-view inconsistency, leading to inferior rendering quality. To address this challenge, we propose a novel Depth-aware Gaussian Splatting for efficient 3D scene Restoration, called RestorGS, which flexibly restores multiple degraded scenes using a unified framework. Specifically, RestorGS consists of two core designs: Appearance Decoupling and Depth-Guided Modeling. The former exploits appearance learning over spherical harmonics to decouple clear and degraded Gaussian, thus separating the clear views from the degraded ones. Collaboratively, the latter leverages the depth information to guide the degradation modeling, thereby facilitating the decoupling process. Benefiting from the above optimization strategy, our method achieves high-quality restoration while enabling real-time rendering speed. Extensive experiments show that our RestorGS outperforms existing methods significantly in underwater, nighttime, and hazy scenes.
****

IM-Portrait: Learning 3D-aware Video Diffusion for Photorealistic Talking Heads from Monocular VideosC
Yuan Li, Ziqian Bai, Feitong Tan, Zhaopeng Cui, Sean Fanello, Yinda Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21107-21116
We propose a novel 3D-aware diffusion-based method for generating photorealistic talking head videos directly from a single identity image and explicit control signals (e.g., expressions). Our method generates Multiplane Images (MPIs) that ensure geometric consistency, making them ideal for immersive viewing experiences like binocular videos for VR headsets.Unlike existing methods that often require a separate stage or joint optimization to reconstruct a 3D representation (such as NeRF or 3D Gaussians), our approach directly generates the final output through a single denoising process, eliminating the need for post-processing steps to render novel views efficiently.To effectively learn from monocular videos, we introduce a training mechanism that reconstructs the output MPI randomly in either the target or the reference camera space. This approach enables the model to simultaneously learn sharp image details and underlying 3D information.Extensive experiments demonstrate the effectiveness of our method, which achieves competitive avatar quality and novel-view rendering capabilities, even without explicit 3D reconstruction or high-quality multi-view training data.
****
Deep Fair Multi-View Clustering with Attention KAN
HaiMing Xu, Qianqian Wang, Boyue Wang, Quanxue Gao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5061-5070
Multi-view clustering is effective in unsupervised multi-view data analysis and has received considerable attention. However, most existing methods excessively emphasize certain attributes, resulting in unfair clustering outcomes, i.e., certain sensitive attributes dominate the clustering results. Moreover, existing methods struggle to effectively capture complex nonlinear relationships and interactions across views, limiting their ability to achieve optimal clustering performance. Therefore, in this work, we propose a novel method, Deep Fair Multi-View Clustering with Attention Kolmogorov-Arnold Network (DFMVC-AKAN), to generate fair clustering results while maintaining robust performance. DFMVC-AKAN integrates attention mechanisms into Kolmogorov-Arnold Networks (KAN) to exploit the complex nonlinear inter-view relationships. Specifically, KAN provides a nonlinear feature representation capable of efficiently approximating arbitrary multivariate continuous functions, augmented by a hybrid attention mechanism which enables the model to dynamically focus on the most relevant features. Finally, we refine the clustering assignments with a distribution alignment module to ensure fair outcomes across diverse groups while maintaining discriminative ability. Experimental results on four datasets containing sensitive attributes demonstrate that DFMVC-AKAN significantly improves fairness and clustering performance compared to state-of-the-art methods.
****
LineArt: A Knowledge-guided Training-free High-quality Appearance Transfer for Design Drawing with Diffusion Model
Xi Wang, Hongzhen Li, Heng Fang, Yichen Peng, Haoran Xie, Xi Yang, Chuntao Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2912-2923
Image rendering from line drawings is vital in design and image generation technologies reduce costs, yet professional line drawings demand preserving complex details. Text prompts struggle with accuracy, and image translation struggles with consistency and fine-grained control. We present LineArt, a framework that transfers complex appearance onto detailed design drawings, facilitating design and artistic creation. It generates high-fidelity appearance while preserving structural accuracy by simulating hierarchical visual cognition and integrating human artistic experience to guide the diffusion process. LineArt overcomes the limitations of current methods in terms of difficulty in fine-grained control and style degradation in design drawings. It requires no precise 3D modeling, physical property specifications, or network training, making it more convenient for desi

gn tasks. LineArt consists of two stages: a multi-frequency lines fusion module to supplement the input design drawing with detailed structural information and a two-part painting process for Base Layer Shaping and Surface Layer Coloring. We also present a new design drawing dataset, ProLines, for evaluation. The experiments show that LineArt performs better in accuracy, realism, and material precision compared to SOTAs.

****

4Real-Video: Learning Generalizable Photo-Realistic 4D Video Diffusion

Chaoyang Wang, Peiye Zhuang, Tuan Duc Ngo, Willi Menapace, Aliaksandr Siarohin, Michael Vasilkovsky, Ivan Skorokhodov, Sergey Tulyakov, Peter Wonka, Hsin-Ying Lee; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17723-17732

We propose 4Real-Video, a novel framework for generating 4D videos, organized as a grid of video frames with both time and viewpoint axes. In this grid, each row contains frames sharing the same timestep, while each column contains frames from the same viewpoint. One stream performs viewpoint updates on columns, and the other stream performs temporal updates on rows. After each diffusion transformer layer, a newly designed synchronization layer exchanges information between the two token streams. We propose two implementations of the synchronization layer, using either hard or soft synchronization.This feedforward architecture improves upon previous work in three ways: higher inference speed, enhanced visual quality (measured by FVD, CLIP, and VideoScore), and improved temporal and viewpoint consistency (measured by VideoScore, GIM-Confidence, and Dust3R-Confidence).

****

DynaMoDe-NeRF: Motion-aware Deblurring Neural Radiance Field for Dynamic Scenes

Ashish Kumar, Rajagopalan A. N.; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21728-21738

Neural Radiance Fields (NeRFs) have made significant advances in rendering novel photorealistic views for both static and dynamic scenes. However, most prior works assume ideal conditions of artifact-free visual inputs i.e., images and videos. In real scenarios, artifacts such as object motion blur, camera motion blur, or lens defocus blur are ubiquitous. Some recent studies have explored novel view synthesis using blurred input frames by examining either camera motion blur, defocus blur, or both. However, these studies are limited to static scenes. In this work, we enable NeRFs to deal with object motion blur whose local nature stems from the interplay between object velocity and camera exposure time. Often, the object motion is unknown and time varying, and this adds to the complexity of scene reconstruction. Sports videos are a prime example of how rapid object motion can significantly degrade video quality for static cameras by introducing motion blur. We present an approach for realizing motion blur-free novel views of dynamic scenes from input videos with object motion blur captured from static cameras spanning multiple poses. We propose a NeRF-based analytical framework that elegantly correlates object three-dimensional (3D) motion across views as well as time to the observed blurry videos. Our proposed method DynaMoDe-NeRF (Dynamic Motion-aware Deblurring NeRF) is self-supervised and reconstructs the dynamic 3D scene, renders sharp novel views by blind deblurring, and recovers the underlying 3D motion and blur parameters. We provide comprehensive experimental analysis on synthetic and real data to validate our approach. To the best of our knowledge, this is the first work to address localized object motion blur in the NeRF domain.

****

VideoICL: Confidence-based Iterative In-context Learning for Out-of-Distribution Video Understanding

Kangsan Kim, Geon Park, Youngwan Lee, Woongyeong Yeo, Sung Ju Hwang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3295-3305

Recent advancements in video large multimodal models (LMMs) have significantly improved their video understanding and reasoning capabilities. However, their performance drops on out-of-distribution (OOD) tasks that are underrepresented in training data. Traditional methods like fine-tuning on OOD datasets are impractic

al due to high computational costs. While In-context learning (ICL) with demonstration examples has shown promising generalization performance in language tasks and image-language tasks without fine-tuning, applying ICL to video-language tasks faces challenges due to the limited context length in Video LMMs, as videos require longer token lengths. To address these issues, we propose VideoICL, a novel video in-context learning framework for OOD tasks that introduces a similarity-based relevant example selection strategy and a confidence-based iterative inference approach. This allows to select the most relevant examples and rank them based on similarity, to be used for inference. If the generated response has low confidence, our framework selects new examples and performs inference again, iteratively refining the results until a high-confidence response is obtained. This approach improves OOD video understanding performance by extending effective context length without incurring high costs. The experimental results on multiple benchmarks demonstrate significant performance gains, especially in domain-specific scenarios, laying the groundwork for broader video comprehension applications.

****

Zero-Shot Image Restoration Using Few-Step Guidance of Consistency Models (and Beyond)

Tomer Garber, Tom Tirer; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2398-2407

In recent years, it has become popular to tackle image restoration tasks with a single pretrained diffusion model (DM) and data-fidelity guidance, instead of training a dedicated deep neural network per task. However, such "zero-shot" restoration schemes currently require many Neural Function Evaluations (NFEs) for performing well, which may be attributed to the many NFEs needed in the original generative functionality of the DMs. Recently, faster variants of DMs have been explored for image generation. These include Consistency Models (CMs), which can generate samples via a couple of NFEs. However, existing works that use guided CMs for restoration still require tens of NFEs or fine-tuning of the model per task that leads to performance drop if the assumptions during the fine-tuning are not accurate. In this paper, we propose a zero-shot restoration scheme that uses CMs and operates well with as little as 4 NFEs. It is based on a wise combination of several ingredients: better initialization, back-projection guidance, and above all a novel noise injection mechanism. We demonstrate the advantages of our approach for image super-resolution and inpainting. Interestingly, we show that the usefulness of our noise injection technique goes beyond CMs: it can also mitigate the performance degradation of existing guided DM methods when reducing their NFE count.

****

Similarity-Guided Layer-Adaptive Vision Transformer for UAV Tracking

Chaocan Xue, Bineng Zhong, Qihua Liang, Yaozong Zheng, Ning Li, Yuanliang Xue, Shuxiang Song; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6730-6740

Vision transformers (ViTs) have emerged as a popular backbone for visual tracking. However, complete ViT architectures are too cumbersome to deploy for unmanned aerial vehicle (UAV) tracking which extremely emphasizes efficiency. In this study, we discover that many layers within lightweight ViT-based trackers tend to learn relatively redundant and repetitive target representations. Based on this observation, we propose a similarity-guided layer adaptation approach to optimize the structure of ViTs. Our approach dynamically disables a large number of representation-similar layers and selectively retains only a single optimal layer among them, aiming to achieve a better accuracy-speed trade-off. By incorporating this approach into existing ViTs, we tailor previously complete ViT architectures into an efficient similarity-guided layer-adaptive framework, namely SGLATrack, for real-time UAV tracking. Extensive experiments on six tracking benchmarks verify the effectiveness of the proposed approach, and show that our SGLATrack achieves a state-of-the-art real-time speed while maintaining competitive tracking precision. Codes and models are available at https://github.com/GXNU-ZhongLab/SGLATrack.

****
LidarGait++: Learning Local Features and Size Awareness from LiDAR Point Clouds for 3D Gait Recognition
Chuanfu Shen, Rui Wang, Lixin Duan, Shiqi Yu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6627-6636
Point clouds have gained growing interest in gait recognition. However, current methods, which typically convert point clouds into 3D voxels, often fail to extract essential gait-specific features. In this paper, we explore gait recognition within 3D point clouds from the perspectives of architectural designs and gait representation modeling. We indicate the significance of local and body size features in 3D gait recognition and introduce LidarGait++, a novel framework combining advanced local representation learning techniques with a novel size-aware learning mechanism. Specifically, LidarGait++ utilizes Set Abstraction (SA) layer and Pyramid Point Pooling (P^3) layer for learning locally fine-grained gait representations from 3D point clouds directly. Both the SA and P^3 layers can be further enhanced with size-aware learning to make the model aware of the actual size of the subjects. In the end, LidarGait++ not only outperforms current state-of-the-art methods, but it also consistently demonstrates robust performance and great generalizability on two benchmarks. Our extensive experiments validate the effectiveness of size and local features in 3D gait recognition.
****
UrbanCAD: Towards Highly Controllable and Photorealistic 3D Vehicles for Urban Scene Simulation
Yichong Lu, Yichi Cai, Shangzhan Zhang, Hongyu Zhou, Haoji Hu, Huimin Yu, Andreas Geiger, Yiyi Liao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27519-27530
Photorealistic 3D vehicle models with high controllability are essential for autonomous driving simulation and data augmentation. While handcrafted CAD models provide flexible controllability, free CAD libraries often lack the high-quality materials necessary for photorealistic rendering. Conversely, reconstructed 3D models offer high-fidelity rendering but lack controllability. In this work, we introduce UrbanCAD, a framework that generates highly controllable and photorealistic 3D vehicle digital twins from a single urban image, leveraging a large collection of free 3D CAD models and handcrafted materials. To achieve this, we propose a novel pipeline that follows a retrieval-optimization manner, adapting to observational data while preserving fine-grained expert-designed priors for both geometry and material. This enables vehicles' realistic 360-degree rendering, background insertion, material transfer, relighting, and component manipulation. Furthermore, given multi-view background perspective and fisheye images, we approximate environment lighting using fisheye images and reconstruct the background with 3DGS, enabling the photorealistic insertion of optimized CAD models into rendered novel view backgrounds. Experimental results demonstrate that UrbanCAD outperforms baselines in terms of photorealism. Additionally, we show that various perception models maintain their accuracy when evaluated on UrbanCAD with in-distribution configurations but degrade when applied to realistic out-of-distribution data generated by our method. This suggests that UrbanCAD is a significant advancement in creating photorealistic, safety-critical driving scenarios for downstream applications.
****
Coarse Correspondences Boost Spatial-Temporal Reasoning in Multimodal Language Model
Benlin Liu, Yuhao Dong, Yiqin Wang, Zixian Ma, Yansong Tang, Luming Tang, Yongming Rao, Wei-Chiu Ma, Ranjay Krishna; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3783-3792
Multimodal language models (MLLMs) are increasingly being applied in real-world environments, necessitating their ability to interpret 3D spaces and comprehend temporal dynamics. Current methods often rely on specialized architectural designs or task-specific fine-tuning to achieve this. We introduce Coarse Correspondences, a simple lightweight method that enhances MLLMs' spatial-temporal reasoning with 2D images as input, without modifying the architecture or requiring task-

specific fine-tuning. Our method uses a lightweight tracking model to identify p
rimary object correspondences between frames in a video or across different imag
e viewpoints, and then conveys this information to MLLMs through visual promptin
g. We demonstrate that this simple training-free approach brings substantial gai
ns to GPT4-V/O consistently on four benchmarks that require spatial-temporal rea
soning, including +20.5% improvement on ScanQA, +9.7% on OpenEQA's episodic memo
ry subset, +6.0% on the long-form video benchmark EgoSchema, and +11% on the R2R
 navigation benchmark. Additionally, we show that Coarse Correspondences can al
so enhance open-source MLLMs' spatial reasoning (by +6.9% on ScanQA) when applie
d in both training and inference and that the improvement can generalize to unse
en datasets such as SQA3D (+3.1%). Taken together, we show that Coarse Correspon
dences effectively and efficiently boosts models' performance on downstream tas
ks requiring spatial-temporal reasoning.
****

Diff-Palm: Realistic Palmprint Generation with Polynomial Creases and Intra-Clas
s Variation Controllable Diffusion Models

Jianlong Jin, Chenglong Zhao, Ruixin Zhang, Sheng Shang, Jianqing Xu, Jingyun Zh
ang, ShaoMing Wang, Yang Zhao, Shouhong Ding, Wei Jia, Yunsheng Wu; Proceedings
of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2636
7-26376

Palmprint recognition is significantly limited by the lack of large-scale public
ly available datasets. Previous methods have adopted Bezier curves to simulate t
he palm creases, which then serve as input for conditional GANs to generate real
istic palmprints.However, without employing real data fine-tuning, the performan
ce of the recognition model trained on these synthetic datasets would drasticall
y decline, indicating a large gap between generated and real palmprints.This is
primarily due to the utilization of an inaccurate palm crease representation and
 challenges in balancing intra-class variation with identity consistency.To addr
ess this, we introduce a polynomial-based palm crease representation that provid
es a new palm crease generation mechanism more closely aligned with the real dis
tribution. We also propose the palm creases conditioned diffusion model with a n
ovel intra-class variation control method.By applying our proposed K-step noise-
sharing sampling, we are able to synthesize palmprint datasets with large intra-
class variation and high identity consistency.Experimental results show that, fo
r the first time, recognition models trained solely on our synthetic datasets, w
ithout any fine-tuning, outperform those trained on real datasets.Furthermore, o
ur approach achieves superior recognition performance as the number of generated
 identities increases.Our code and pre-trained models will be released.
****

FoundationStereo: Zero-Shot Stereo Matching

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, Stan Birchfi
eld; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR
), 2025, pp. 5249-5260

Tremendous progress has been made in deep stereo matching to excel on benchmark
datasets through per-domain fine-tuning. However, achieving strong zero-shot gen
eralization - a hallmark of foundation models in other computer vision tasks - r
emains challenging for stereo matching. We introduce FoundationStereo, a foundat
ion model for stereo depth estimation designed to achieve strong zero shot gener
alization. To this end, we first construct a large scale (1M stereo pairs) synth
etic training dataset featuring large diversity and high photorealism, followed
by an automatic self-curation pipeline to remove ambiguous samples. We then desi
gn a number of network architecture components to enhance scalability, including
 a side-tuning feature backbone that adapts rich monocular priors from vision fo
undation models to mitigate the sim-to-real gap, and long-range context reasonin
g for effective cost volume filtering. Together, these components lead to strong
 robustness and accuracy across domains, establishing a new standard in zero-sho
t stereo depth estimation. Project page: https://nvlabs.github.io/FoundationSter
eo
****

Z-Magic: Zero-shot Multiple Attributes Guided Image Creator

Yingying Deng, Xiangyu He, Fan Tang, Weiming Dong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18390-18400
The customization of multiple attributes has gained increasing popularity with the rising demand for personalized content creation. Despite promising empirical results, the contextual coherence between different attributes has been largely overlooked. In this paper, we argue that subsequent attributes should follow the multivariable conditional distribution introduced by former attributes creation. In light of this, we reformulate multi-attribute creation from a conditional probability theory perspective and tackle the challenging zero-shot setting. By explicitly modeling the dependencies between attributes, we further enhance the coherence of generated images across diverse attribute combinations. Furthermore, we identify connections between multi-attribute customization and multi-task learning, effectively addressing the high computing cost encountered in multi-attribute synthesis. Extensive experiments demonstrate that Z-Magic outperforms existing models in zero-shot image generation, with broad implications for AI-driven design and creative applications.
****

UniNet: A Contrastive Learning-guided Unified Framework with Feature Selection for Anomaly Detection
Shun Wei, Jielin Jiang, Xiaolong Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9994-10003
Anomaly detection (AD) is a crucial visual task aimed at recognizing abnormal pattern within samples. However, most existing AD methods suffer from limited generalizability, as they are primarily designed for domain-specific applications, such as industrial scenarios, and often perform poorly when applied to other domains. This challenge largely stems from the inherent discrepancies in features across domains. To bridge this domain gap, we introduce UniNet, a generic unified framework that incorporates effective feature selection and contrastive learning-guided anomaly discrimination. UniNet comprises student-teacher models and a bottleneck, featuring several vital innovations: First, we propose domain-related feature selection, where the student is guided to select and focus on representative features from the teacher with domain-relevant priors, while restoring them effectively. Second, a similarity contrastive loss function is developed to strengthen the correlations among homogeneous features. Meanwhile, a margin loss function is proposed to enforce the separation between the similarities of abnormality and normality, effectively improving the model's ability to discriminate anomalies. Third, we propose a weighted decision mechanism for dynamically evaluating the anomaly score to achieve robust AD. Large-scale experiments on 12 datasets from various domains show that UniNet surpasses existing methods.
****

Chain of Semantics Programming in 3D Gaussian Splatting Representation for 3D Vision Grounding
Jiaxin Shi, Mingyue Xiang, Hao Sun, Yixuan Huang, Zhi Weng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24560-24569
3D Vision Grounding (3DVG) is a fundamental research area that enables agents to perceive and interact with the 3D world. The challenge of the 3DVG task lies in understanding fine-grained semantics and spatial relationships within both the utterance and 3D scene. To address this challenge, we propose a zero-shot neuro-symbolic framework that utilizes a large language model (LLM) as neuro-symbolic functions to ground the object within the 3D Gaussian Splatting (3DGS) representation. By utilizing 3DGS representation, we can dynamically render high-quality 2D images from various viewpoints to enrich the semantic information. Given the complexity of spatial relationships, we construct a relationship graph and chain of semantics that decouple spatial relationships and facilitate step-by-step reasoning within 3DGS representation. Additionally, we employ a grounded-aware self-check mechanism to enable the LLM to reflect on its responses and mitigate the effects of ambiguity in spatial reasoning. We evaluate our method using two publicly available datasets, Nr3D and Sr3D, achieving accuracies of 60.8% and 91.4%, respectively. Notably, our method surpasses current state-of-the-art zero-shot methods on the Nr3D dataset. In addition, it outperforms the recent supervised

models on the Sr3D dataset.
****
On the Zero-shot Adversarial Robustness of Vision-Language Models: A Truly Zero-shot and Training-free Approach
Baoshun Tong, Hanjiang Lai, Yan Pan, Jian Yin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19921-19930
Pre-trained Vision-Language Models (VLMs) like CLIP, have demonstrated strong zero-shot generalization capabilities. Despite their effectiveness on various downstream tasks, they remain vulnerable to adversarial samples. Existing methods fine-tune VLMs to improve their performance via performing adversarial training on a certain dataset. However, this can lead to model overfitting and is not a true zero-shot scenario. In this paper, we propose a truly zero-shot and training-free approach that can significantly improve the VLM's zero-shot adversarial robustness. Specifically, we first discover that simply adding Gaussian noise greatly enhances the VLM's zero-shot performance. Then, we treat the adversarial examples with added Gaussian noise as anchors and strive to find a path in the embedding space that leads from the adversarial examples to the cleaner samples. We improve the VLMs' generalization abilities in a truly zero-shot and training-free manner compared to previous methods. Extensive experiments on 16 datasets demonstrate that our method can achieve state-of-the-art zero-shot robust performance, improving the top-1 robust accuracy by an average of 9.77%. The code will be publicly available.
****
Towards General Visual-Linguistic Face Forgery Detection
Ke Sun, Shen Chen, Taiping Yao, Ziyin Zhou, Jiayi Ji, Xiaoshuai Sun, Chia-Wen Lin, Rongrong Ji; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19576-19586
Face manipulation techniques have achieved significant advances, presenting serious challenges to security and social trust. Recent works demonstrate that leveraging multimodal models can enhance the generalization and interpretability of face forgery detection. However, existing annotation approaches, whether through human labeling or direct Multimodal Large Language Model (MLLM) generation, often suffer from hallucination issues, leading to inaccurate text descriptions, especially for high-quality forgeries. To address this, we propose Face Forgery Text Generator (FFTG), a novel annotation pipeline that generates accurate text descriptions by leveraging forgery masks for initial region and type identification, followed by a comprehensive prompting strategy to guide MLLMs in reducing hallucination. We validate our approach through fine-tuning both CLIP with a three-branch training framework combining unimodal and multimodal objectives, and MLLMs with our structured annotations. Experimental results demonstrate that our method not only achieves more accurate annotations with higher region identification accuracy, but also leads to improvements in model performance across various forgery detection benchmarks.
****
Movie Weaver: Tuning-Free Multi-Concept Video Personalization with Anchored Prompts
Feng Liang, Haoyu Ma, Zecheng He, Tingbo Hou, Ji Hou, Kunpeng Li, Xiaoliang Dai, Felix Juefei-Xu, Samaneh Azadi, Animesh Sinha, Peizhao Zhang, Peter Vajda, Diana Marculescu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13146-13156
Video personalization, which generates customized videos using reference images, has gained significant attention.However, prior methods typically focus on single-concept personalization, limiting broader applications that require multi-concept integration.Attempts to extend these models to multiple concepts often lead to identity blending, which results in composite characters with fused attributes from multiple sources.This challenge arises due to the lack of a mechanism to link each concept with its specific reference image.We address this with anchored prompts, which embed image anchors as unique tokens within text prompts, guiding accurate referencing during generation.Additionally, we introduce concept embeddings to encode the order of reference images.Our approach, Movie Weaver, sea

mlessly weaves multiple concepts--including face, body, and animal images--into one video, allowing flexible combinations in a single model.The evaluation shows that Movie Weaver outperforms existing methods for multi-concept video personalization in identity preservation and overall quality.
****

LongVALE: Vision-Audio-Language-Event Benchmark Towards Time-Aware Omni-Modal Perception of Long Videos
Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, Feng Zheng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18959-18969
Despite impressive advancements in video understanding, most efforts remain limited to coarse-grained or visual-only video tasks. However, real-world videos encompass omni-modal information (vision, audio, and speech) with a series of events forming a cohesive storyline. The lack of multi-modal video data with fine-grained event annotations and the high cost of manual labeling are major obstacles to comprehensive omni-modality video perception. To address this gap, we propose an automatic pipeline consisting of high-quality multi-modal video filtering, semantically coherent omni-modal event boundary detection, and cross-modal correlation-aware event captioning. In this way, we present LongVALE, the first-ever Vision-Audio-Language Event understanding benchmark comprising 105K omni-modal events with precise temporal boundaries and detailed relation-aware captions within 8.4K high-quality long videos. Further, we build a baseline that leverages LongVALE to enable video large language models (LLMs) for omni-modality fine-grained temporal video understanding for the first time. Extensive experiments demonstrate the effectiveness and great potential of LongVALE in advancing comprehensive multi-modal video understanding.
****

MVPortrait: Text-Guided Motion and Emotion Control for Multi-view Vivid Portrait Animation
Yukang Lin, Hokit Fung, Jianjin Xu, Zeping Ren, Adela S.M. Lau, Guosheng Yin, Xiu Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26242-26252
Recent portrait animation methods have made significant strides in generating realistic lip synchronization. However, they often lack explicit control over head movements and facial expressions, and cannot produce videos from multiple viewpoints, resulting in less controllable and expressive animations. Moreover, text-guided portrait animation remains underexplored, despite its user-friendly nature. In this paper, we present a novel two-stage text-guided framework, MVPortrait, to generate expressive multi-view portrait animations that faithfully capture the described motion and emotion. MVPortrait is the first to introduce FLAME as an intermediate representation, effectively embedding facial movements, expressions, and view transformations within its parameter space. In the first stage, we separately train the FLAME motion and emotion diffusion models based on text input. In the second stage, we train a multi-view video generation model conditioned on a reference portrait image and multi-view FLAME rendering sequences from the first stage. Experimental results exhibit that MVPortrait outperforms existing methods in terms of motion and emotion control, as well as view consistency. Furthermore, by leveraging FLAME as a bridge, MVPortrait becomes the first controllable portrait animation framework that is compatible with text, speech, and video as driving signals.
****

MoEdit: On Learning Quantity Perception for Multi-object Image Editing
Yanfeng Li, Kahou Chan, Yue Sun, Chantong Lam, Tong Tong, Zitong Yu, Keren Fu, Xiaohong Liu, Tao Tan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2683-2693
Multi-object images are widely present in the real world, spanning various areas of daily life. Efficient and accurate editing of these images is crucial for applications such as augmented reality, advertisement design, and medical imaging. Stable Diffusion (SD) has ushered in a new era of high-quality image generation and editing. However, existing methods struggle to analyze abstract relationshi

ps between multiple objects, often yielding suboptimal performance. To address t
his, we propose MoEdit, an auxiliary-free method for multi-object image editing.
 This method enables high-quality editing of attributes in multi-object images,
such as style, object features, and background, by maintaining quantity consiste
ncy between the input and output images. To preserve the concept of quantity, we
 introduce a Quantity Attention (QTTN) module to control the editing process. Ad
ditionally, we present a Feature Compensation (FeCom) module to enhance the robu
stness of object preservation. We also employ a null-text training strategy to r
etain the guiding capability of text prompts, making them plug-and-play modules.
 This method leverages the powerful image perception and generation capabilities
 of the SD model, enabling specific concepts to be preserved and altered between
 input and output images. Experimental results demonstrate that MoEdit achieves
the state-of-the-art performance in high-quality multi-object image editing. Dat
a and codes will be available at https://github.com/Tear-kitty/MoEdit.
****

Seeing More with Less: Human-like Representations in Vision Models
Andrey Gizdov, Shimon Ullman, Daniel Harari; Proceedings of the Computer Vision
and Pattern Recognition Conference (CVPR), 2025, pp. 4408-4417
Large multimodal models (LMMs) typically process visual inputs with uniform reso
lution across the entire field of view, leading to inefficiencies when non-criti
cal image regions are processed as precisely as key areas. Inspired by the human
 visual system's foveated approach, we apply a sampling method to leading archit
ectures such as MDETR, BLIP2, InstructBLIP, LLaVA, and ViLT, and evaluate their
performance with variable (foveated) resolution inputs. Results show that foveat
ed sampling boosts accuracy in visual tasks like question answering and object d
etection under tight pixel budgets, improving performance by up to 2.7% on the G
QA dataset, 2.1% on SEED-Bench, and 2.0% on VQAv2 compared to uniform sampling.
Furthermore, we show that indiscriminate resolution increases yield diminishing
returns, with models achieving up to 80% of their full capability using just 3%
of the pixels, even on complex tasks. Foveated sampling prompts more human-like
processing within models, such as neuronal selectivity and globally acting self-
attention in vision transformers. This paper provides a foundational analysis of
 foveated sampling's impact on existing models, suggesting that more efficient a
rchitectural adaptations, mimicking human visual processing, are a promising res
earch venue for the community. Potential applications of our findings center low
 power minimal bandwidth devices (such as UAVs and edge devices), where compact
and efficient vision is critical.
****

Modeling Thousands of Human Annotators for Generalizable Text-to-Image Person Re
-identification
Jiayu Jiang, Changxing Ding, Wentao Tan, Junhong Wang, Jin Tao, Xiangmin Xu; Pro
ceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025,
 pp. 9220-9230
Text-to-image person re-identification (ReID) aims to retrieve the images of an
interested person based on textual descriptions. One main challenge for this tas
k is the high cost in manually annotating large-scale databases, which affects t
he generalization ability of ReID models. Recent works handle this problem by le
veraging Multi-modal Large Language Models (MLLMs) to describe pedestrian images
 automatically. However, the captions produced by MLLMs lack diversity in descri
ption styles. To address this issue, we propose a Human Annotator Modeling (HAM)
 approach to enable MLLMs to mimic the description styles of thousands of human
annotators. Specifically, we first extract style features from human textual des
criptions and perform clustering on them. This allows us to group textual descri
ptions with similar styles into the same cluster. Then, we employ a prompt to re
present each of these clusters and apply prompt learning to mimic the descriptio
n styles of different human annotators. Furthermore, we define a style feature s
pace and perform uniform sampling in this space to obtain more diverse clusterin
g prototypes, which further enriches the diversity of the MLLM-generated caption
s. Finally, we adopt HAM to automatically annotate a massive-scale database for
text-to-image ReID. Extensive experiments on this database demonstrate that it s

ignificantly improves the generalization ability of ReID models.
****

Accelerating Multimodal Large Language Models by Searching Optimal Vision Token Reduction
Shiyu Zhao, Zhenting Wang, Felix Juefei-Xu, Xide Xia, Miao Liu, Xiaofang Wang, Mingfu Liang, Ning Zhang, Dimitris N. Metaxas, Licheng Yu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29869-29879
Prevailing Multimodal Large Language Models (MLLMs) encode the input image(s) as vision tokens and feed them into the language backbone, similar to how Large Language Models (LLMs) process the text tokens. However, the number of vision tokens increases quadratically as the image resolutions, leading to huge computational costs.In this paper, we consider improving MLLM's efficiency from two scenarios, (I) Reducing computational cost without degrading the performance. (II) Improving the performance with given budgets. We start with our main finding that the ranking of each vision token sorted by attention scores is similar in each layer except the first layer. Based on it, we assume that the number of essential top vision tokens does not increase along layers. Accordingly, for Scenario I, we propose a greedy search algorithm (G-Search) to find the least number of vision tokens to keep at each layer from the shallow to the deep. Interestingly, G-Search is able to reach the optimal reduction strategy based on our assumption. For Scenario II, based on the reduction strategy from G-Search, we design a parametric sigmoid function (P-Sigmoid) to guide the reduction at each layer of the MLLM, whose parameters are optimized by Bayesian Optimization. Extensive experiments demonstrate that our approach can significantly accelerate those popular MLLMs, e.g. LLaVA, and InternVL2 models, by more than 2 xwithout performance drops. Our approach also far outperforms other token reduction methods when budgets are limited, achieving a better trade-off between efficiency and effectiveness.
****

Matrix-Free Shared Intrinsics Bundle Adjustment
Daniel Safari; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27017-27026
Research on accelerating bundle adjustment has focused on photo collections where each image is accompanied by its own set of camera parameters. However, real-world applications overwhelmingly call for shared intrinsics bundle adjustment (SI-BA) where camera parameters are shared across multiple images. Utilizing overlooked optimization opportunities specific to SI-BA, most notably matrix-free computation, we present a solver that is eight times faster than alternatives while consuming a tenth of the memory. Additionally, we examine factors contributing to BA instability under single-precision computation and propose mitigations.
****

AeroGen: Enhancing Remote Sensing Object Detection with Diffusion-Driven Data Generation
Datao Tang, Xiangyong Cao, Xuan Wu, Jialin Li, Jing Yao, Xueru Bai, Dongsheng Jiang, Yin Li, Deyu Meng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3614-3624
Remote sensing image object detection (RSIOD) aims to identify and locate specific objects within satellite or aerial imagery. However, there is a scarcity of labeled data in current RSIOD datasets, which significantly limits the performance of current detection algorithms. Although existing techniques, e.g., data augmentation and semi-supervised learning, can mitigate this scarcity issue to some extent, they are heavily dependent on high-quality labeled data and perform worse in rare object classes. To address this issue, this paper proposes a layout-controllable diffusion generative model (i.e. AeroGen) tailored for RSIOD. To our knowledge, AeroGen is the first model to simultaneously support horizontal and rotated bounding box condition generation, thus enabling the generation of high-quality synthetic images that meet specific layout and object category requirements. Additionally, we propose an end-to-end data augmentation framework that integrates a diversity-conditioned generator and a filtering mechanism to enhance both the diversity and quality of generated data. Experimental results demonstrate that the synthetic data produced by our method are of high quality and diversit

y. Furthermore, the synthetic RSIOD data can significantly improve the detection performance of existing RSIOD models, i.e., the mAP metrics on DIOR, DIOR-R, and HRSC datasets are improved by 3.7%, 4.3%, and 2.43%, respectively.
****

Tra-MoE: Learning Trajectory Prediction Model from Multiple Domains for Adaptive Policy Conditioning
Jiange Yang, Haoyi Zhu, Yating Wang, Gangshan Wu, Tong He, Limin Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6960-6970
Learning from multiple domains is a primary factor that influences the generalization of a single unified robot system. In this paper, we aim to learn the trajectory prediction model by using broad out-of-domain data to improve its performance and generalization ability. Trajectory model is designed to predict any-point trajectories in the current frame given an instruction and can provide detailed control guidance for robotic policy learning. To handle the diverse out-of-domain data distribution, we propose a sparsely-gated MoE (Top-1 gating strategy) architecture for trajectory model, coined as Tra-MoE. The sparse activation design enables good balance between parameter cooperation and specialization, effectively benefiting from large-scale out-of-domain data while maintaining constant FLOPs per token. In addition, we further introduce an adaptive policy conditioning technique by learning 2D mask representations for predicted trajectories, which is explicitly aligned with image observations to guide action prediction more flexibly. We perform extensive experiments on both simulation and real-world scenarios to verify the effectiveness of Tra-MoE and adaptive policy conditioning technique. We also conduct a comprehensive empirical study to train Tra-MoE, demonstrating that our Tra-MoE consistently exhibits superior performance compared to the dense baseline model, even when the latter is scaled to match Tra-MoE's parameter count.
****

Mitigating Hallucinations in Large Vision-Language Models via DPO: On-Policy Data Hold the Key
Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, Dongsheng Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10610-10620
Hallucination remains a major challenge for Large Vision-Language Models (LVLMs). Direct Preference Optimization (DPO) has gained increasing attention as a simple solution to hallucination issues. It directly learns from constructed preference pairs that reflect the severity of hallucinations in responses to the same prompt and image. Nonetheless, different data construction methods in existing works bring notable performance variations. We identify a crucial factor here: outcomes are largely contingent on whether the constructed data aligns on-policy w.r.t the initial (reference) policy of DPO. Theoretical analysis suggests that learning from off-policy data is impeded by the presence of KL-divergence between the updated policy and the reference policy. From the perspective of dataset distribution, we systematically summarize the inherent flaws in existing algorithms that employ DPO to address hallucination issues. To alleviate the problems, we propose On-Policy Alignment (OPA)-DPO framework, which uniquely leverages expert feedback to correct hallucinated responses and aligns both the original and expert-revised responses in an on-policy manner. Notably, with only 4.8k data, OPA-DPO achieves an additional reduction in the hallucination rate of LLaVA-1.5-7B: 13.26% on the AMBER benchmark and 5.39% on the Object-Hal benchmark, compared to the previous SOTA algorithm trained with 16k samples.
****

Style Quantization for Data-Efficient GAN Training
Jian Wang, Xin Lan, Jizhe Zhou, Yuxin Tian, Jiancheng Lv; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7696-7706
Under limited data setting, GANs often struggle to navigate and effectively exploit the input latent space. Consequently, images generated from adjacent variables in a sparse input latent space may exhibit significant discrepancies in realism, leading to suboptimal consistency regularization (CR) outcomes. To address t

his, we propose SQ-GAN, a novel approach that enhances CR by introducing a style space quantization scheme. This method transforms the sparse, continuous input latent space into a compact, structured discrete proxy space, allowing each element to correspond to a specific data point, thereby improving CR performance. Instead of direct quantization, we first map the input latent variables into a less entangled "style" space and apply quantization using a learnable codebook. This enables each quantized code to control distinct factors of variation. Additionally, we optimize the optimal transport distance to align the codebook codes with features extracted from the training data by a foundation model, embedding external knowledge into the codebook and establishing a semantically rich vocabulary that properly describes the training dataset. Extensive experiments demonstrate significant improvements in both discriminator robustness and generation quality with our method.

****

Localizing Events in Videos with Multimodal Queries

Gengyuan Zhang, Mang Ling Ada Fok, Jialu Ma, Yan Xia, Daniel Cremers, Philip Torr, Volker Tresp, Jindong Gu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3339-3351

Localizing events in videos based on semantic queries is a pivotal task in video understanding research and user-oriented applications like video search. Yet, current research predominantly relies on natural language queries (NLQs), overlooking the potential of using multimodal queries (MQs) that incorporate images to flexibly represent semantic queries, particularly when it is difficult to express non-verbal or unfamiliar concepts in words. To bridge this gap, we introduce ICQ, a new benchmark designed for localizing events in videos with MQs, alongside an evaluation dataset ICQ-Highlight. To adapt and reevaluate existing video localization models for this new task, we propose 3 Multimodal Query Adaptation methods and a novel Surrogate Fine-tuning strategy, serving as strong baseline methods. ICQ systematically benchmarks 12 state-of-the-art backbone models, spanning from specialized video localization models to Video Large Language Models. Our extensive experiments highlight the high potential of using MQs in real-world applications. We believe this is a first step toward video event localization with MQs.

****

PhysVLM: Enabling Visual Language Models to Understand Robotic Physical Reachability

Weijie Zhou, Manli Tao, Chaoyang Zhao, Haiyun Guo, Honghui Dong, Ming Tang, Jinqiao Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6940-6949

Understanding the environment and a robot's physical reachability is crucial for task execution. While state-of-the-art vision-language models (VLMs) excel in environmental perception, they often generate inaccurate or impractical responses in embodied visual reasoning tasks due to a lack of understanding of robotic physical reachability. To address this issue, we propose a unified representation of physical reachability across diverse robots, i.e., Space-Physical Reachability Map (S-P Map), and PhysVLM, a vision-language model that integrates this reachability information into visual reasoning. Specifically, the S-P Map abstracts a robot's physical reachability into a generalized spatial representation, independent of specific robot configurations, allowing the model to focus on reachability features rather than robot-specific parameters. Subsequently, PhysVLM extends traditional VLM architectures by incorporating an additional feature encoder to process the S-P Map, enabling the model to reason about physical reachability without compromising its general vision-language capabilities. To train and evaluate PhysVLM, we constructed a large-scale multi-robot dataset, Phys100K, and a challenging benchmark, EQA-phys, which includes tasks for six different robots in both simulated and real-world environments. Experimental results demonstrate that PhysVLM outperforms existing models, achieving a 14% improvement over GPT-4o on EQA-phys and surpassing advanced embodied VLMs such as RoboMamba and SpatialVLM on the RoboVQA-val and OpenEQA benchmarks. Additionally, the S-P Map shows strong compatibility with various VLMs, and its integration into GPT-4o-mini yiel

ds a 7.1% performance improvement.
****
CleanDIFT: Diffusion Features without Noise
Nick Stracke, Stefan Andreas Baumann, Kolja Bauer, Frank Fundel, Björn Ommer; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 117-127
Internal features from large-scale pre-trained diffusion models have recently been established as powerful semantic descriptors for a wide range of downstream tasks. Works that use these features generally need to add noise to images before passing them through the model to obtain the semantic features, as the models do not offer the most useful features when given images with little to no noise. We show that this noise has a critical impact on the usefulness of these features that cannot be remedied by ensembling with different random noises. We address this issue by introducing a lightweight, unsupervised fine-tuning method that enables diffusion backbones to provide high-quality, noise-free semantic features. We show that these features readily outperform previous diffusion features by a wide margin in a wide variety of extraction setups and downstream tasks, offering better performance than even ensemble-based methods at a fraction of the cost.
****
Simpler Diffusion: 1.5 FID on ImageNet512 with Pixel-space Diffusion
Emiel Hoogeboom, Thomas Mensink, Jonathan Heek, Kay Lamerigts, Ruiqi Gao, Tim Salimans; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18062-18071
Latent diffusion models have become the popular choice for scaling up diffusion models for high resolution image synthesis. Compared to pixel-space models that are trained end-to-end, latent models are perceived to be more efficient and to produce higher image quality at high resolution. Here we challenge these notions, and show that pixel-space models can be very competitive to latent models both in quality and efficiency, achieving 1.5 FID on ImageNet512 and new SOTA results on ImageNet128, ImageNet256 and Kinetics600. We present a simple recipe for scaling end-to-end pixel-space diffusion models to high resolutions. 1: Use the sigmoid loss-weighting (Kingma & Gao, 2023) with our prescribed hyper-parameters. 2: Use our simplified memory-efficient architecture with fewer skip-connections. 3: Scale the model to favor processing the image at a high resolution with fewer parameters, rather than using more parameters at a lower resolution. Combining these with guidance intervals, we obtain a family of pixel-space diffusion models we call Simpler Diffusion (SiD2).
****
Uncertainty-Instructed Structure Injection for Generalizable HD Map Construction
Xiaolu Liu, Ruizi Yang, Song Wang, Wentong Li, Junbo Chen, Jianke Zhu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22359-22368
Reliable high-definition (HD) map construction is crucial for the driving safety of autonomous vehicles. While recent studies demonstrate improved performance, their generalization capability across unfamiliar driving scenes remains unexplored. To tackle this issue, we propose UIGenMap, an uncertainty-instructed structure injection approach for generalizable HD map vectorization, which concerns the uncertainty resampling in statistical distribution and employs explicit instance features to reduce the excessive reliance on training data. Specifically, we introduce the perspective-view (PV) detection branch to obtain explicit structural features, in which the uncertainty-aware decoder is designed to dynamically sample probability distributions considering the difference in scenes. With probabilistic embedding and selection, UI2DPrompt is proposed to construct PV learnable prompts. These PV prompts are integrated into the map decoder by designed hybrid injection to compensate for neglected instance structures. To ensure real-time inference, a lightweight Mimic Query Distillation is designed to learn from PV prompts, which can serve as an efficient alternative to the flow of PV branches. Extensive experiments on challenging geographically disjoint (geo-based) data splits demonstrate that our UIGenMap achieves superior performance, with +5.7 m

AP improvement on nuScenes dataset. Source code is available at \href https://github.com/xiaolul2/UIGenMap https://github.com/xiaolul2/UIGenMap .
****
STING-BEE: Towards Vision-Language Model for Real-World X-ray Baggage Security Inspection
Divya Velayudhan, Abdelfatah Ahmed, Mohamad Alansari, Neha Gour, Abderaouf Behouch, Taimur Hassan, Syed Talal Wasim, Nabil Maalej, Muzammal Naseer, Juergen Gall, Mohammed Bennamoun, Ernesto Damiani, Naoufel Werghi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20767-20777
Advancements in Computer-Aided Screening (CAS) systems are essential for improving the detection of security threats in X-ray baggage scans. However, current datasets are limited in representing real-world, sophisticated threats and concealment tactics, and existing approaches are constrained by a closed-set paradigm with predefined labels. To address these challenges, we introduce STCray, the first multimodal X-ray baggage security dataset, comprising 46,642 image-caption paired scans across 21 threat categories, generated using an X-ray scanner for airport security. STCray is meticulously developed with our specialized protocol that ensures domain-aware, coherent captions, that lead to the multi-modal instruction following data in X-ray baggage security. This allows us to train a domain-aware visual AI assistant named STING-BEE that supports a range of vision-language tasks, including scene comprehension, referring threat localization, visual grounding, and visual question answering (VQA), establishing novel baselines for multi-modal learning in X-ray baggage security. Further, STING-BEE shows state-of-the-art generalization in cross-domain settings. Code, data, and models are available at https://divs1159.github.io/STING-BEE/.
****
Not All Parameters Matter: Masking Diffusion Models for Enhancing Generation Ability
Lei Wang, Senmao Li, Fei Yang, Jianye Wang, Ziheng Zhang, Yuhan Liu, Yaxing Wang, Jian Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12880-12890
The diffusion models, in early stages focus on constructing basic image structures, while the refined details, including local features and textures, are generated in later stages. Thus the same network layers are forced to learn both structural and textural information simultaneously, significantly differing from the traditional deep learning architectures (e.g., ResNet or GANs) which captures or generates the image semantic information at different layers. This difference inspires us to explore the time-wise diffusion models. We initially investigate the key contributions of the U-Net parameters to the denoising process and identify that properly zeroing out certain parameters (including large parameters) contributes to denoising, substantially improving the generation quality on the fly. Capitalizing on this discovery, we propose a simple yet effective method--termed "MaskUNet"-- that enhances generation quality with negligible parameter numbers. Our method fully leverages timestep- and sample-dependent effective U-Net parameters. To optimize MaskUNet, we offer two fine-tuning strategies: a training-based approach and a training-free approach, including tailored networks and optimization functions. In zero-shot inference on the COCO dataset, MaskUNet achieves the best FID score and further demonstrates its effectiveness in downstream task evaluations. Project page: https://gudaochangsheng.github.io/MaskUnet-Page/
****
MAD: Memory-Augmented Detection of 3D Objects
Ben Agro, Sergio Casas, Patrick Wang, Thomas Gilles, Raquel Urtasun; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1449-1460
To perceive, humans use memory to fill in gaps caused by our limited visibility, whether due to occlusion or our narrow field of view. However, most 3D object detectors are limited to using sensor evidence from a short temporal window (0.1s-0.3s). In this work, we present a simple and effective add-on for enhancing any existing 3D object detector with long-term memory regardless of its sensor moda

lity (e.g., LiDAR, camera) and network architecture. We propose a model to effec tively align and fuse object proposals from a detector with object proposals fro m a memory bank of past predictions, exploiting trajectory forecasts to align pr oposals across time. We propose a novel schedule to train our model on temporal data that balances data diversity and the gap between training and inference. By applying our method to existing LiDAR and camera-based detectors on the Waymo O pen Dataset (WOD) and Argoverse 2 Sensor (AV2) dataset, we demonstrate significa nt improvements in detection performance (+2.5 to +7.6 AP points). Our method at tains the best performance on the WOD 3D detection leaderboard among online meth ods (excluding ensembles or test-time augmentation).
****

Doppelgangers and Adversarial Vulnerability
George Kamberov; Proceedings of the Computer Vision and Pattern Recognition Conf erence (CVPR), 2025, pp. 10244-10254
Many machine learning (ML) classifiers are claimed to outperform humans, but the y still make mistakes that humans do not. The most notorious examples of such mi stakes are adversarial visual metamers. This paper aims to define and investigat e the phenomenon of adversarial Doppelgangers (AD), which includes adversarial v isual metamers, and to compare the performance and robustness of ML classifiers to human performance.We find that AD are inputs that are close to each other wit h respect to a perceptual metric defined in this paper, and show that AD are qua litatively different from the usual adversarial examples. The vast majority of c lassifiers are vulnerable to AD and robustness-accuracy trade-offs may not impro ve them. Some classification problems may not admit any AD robust classifiers be cause the underlying classes are ambiguous. We provide criteria that can be use d to determine whether a classification problem is well defined or not; describe of an AD robust classifiers' structure and attributes; introduce and explore th e notions of conceptual entropy and regions of conceptual ambiguity for classifi ers that are vulnerable to AD attacks, along with methods to bound the AD foolin g rate of an attack. We define the notion of classifiers that exhibit hyper-sens itive behavior, that is, classifiers whose only mistakes are adversarial Doppelg angers. Improving the AD robustness of hyper-sensitive classifiers proving accur acy. We identify conditions guaranteeing that all classifiers with sufficiently high accuracy are hyper-sensitive.
****

Complexity Experts are Task-Discriminative Learners for Any Image Restoration
Eduard Zamfir, Zongwei Wu, Nancy Mehta, Yuedong Tan, Danda Pani Paudel, Yulun Zh ang, Radu Timofte; Proceedings of the Computer Vision and Pattern Recognition Co nference (CVPR), 2025, pp. 12753-12763
Recent advancements in all-in-one image restoration models have revolutionized t he ability to address diverse degradations through a unified framework. However, parameters tied to specific tasks often remain inactive for other tasks, making mixture-of-experts (MoE) architectures a natural extension. Despite this, MoEs often show inconsistent behavior, with some experts unexpectedly generalizing ac ross tasks while others struggle within their intended scope. This hinders lever aging MoEs' computational benefits by bypassing irrelevant experts during infere nce.We attribute this undesired behavior to the uniform and rigid architecture o f traditional MoEs. To address this, we introduce "complexity experts" -- flexib le expert blocks with varying computational complexity and receptive fields. A k ey challenge is assigning tasks to each expert, as degradation complexity is unk nown in advance. Thus, we execute tasks with a simple bias toward lower complexi ty.To our surprise, this preference effectively drives task-specific allocation, assigning tasks to experts with the appropriate complexity. Extensive experimen ts validate our approach, demonstrating the ability to bypass irrelevant experts during inference while maintaining superior performance. The proposed MoCE-IR m odel outperforms state-of-the-art methods, affirming its efficiency and practica l applicability. The source code and models are publicly available at \href http s://eduardzamfir.github.io/moceir/ \texttt eduardzamfir.github.io/MoCE-IR/
****

Generative Omnimatte: Learning to Decompose Video into Layers

Yao-Chih Lee, Erika Lu, Sarah Rumbley, Michal Geyer, Jia-Bin Huang, Tali Dekel, Forrester Cole; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12522-12532

Given a video and a set of input object masks, an omnimatte method aims to decompose the video into semantically meaningful layers containing individual objects along with their associated effects, such as shadows and reflections.Existing omnimatte methods assume a static background or accurate pose and depth estimation and produce poor decompositions when these assumptions are violated. Furthermore, due to the lack of generative prior on natural videos, existing methods cannot complete dynamic occluded regions.We present a novel generative layered video decomposition framework to address the omnimatte problem. Our method does not assume a stationary scene or require camera pose or depth information and produces clean, complete layers, including convincing completions of occluded dynamic regions. Our core idea is to train a video diffusion model to identify and remove scene effects caused by a specific object. We show that this model can be finetuned from an existing video inpainting model with a small, carefully curated dataset, anddemonstrate high-quality decompositions and editing results for a wide range of casually captured videos containing soft shadows, glossy reflections, splashing water, and more.

****

5%>100%: Breaking Performance Shackles of Full Fine-Tuning on Visual Recognition Tasks

Dongshuo Yin, Leiyi Hu, Bin Li, Youqun Zhang, Xue Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20071-20081

Pre-training & fine-tuning can enhance the transferring efficiency and performance in visual tasks. Recent delta-tuning methods provide more options for visual classification tasks. Despite their success, existing visual delta-tuning art fails to exceed the upper limit of full fine-tuning on challenging tasks. To find a competitive alternative to full fine-tuning, we propose the Multi-cognitive Visual Adapter (Mona) tuning, a novel adapter-based tuning method. First, we introduce multiple vision-friendly filters into the adapter to enhance its ability for processing visual signals, while previous methods mainly rely on language-friendly linear filters. Second, we add the scaled layernorm in the adapter to regulate the distribution of input features for visual filters. To fully demonstrate the practicality and generality of Mona, we conduct experiments on representative visual tasks, including instance segmentation on COCO, semantic segmentation on ADE20K, object detection on Pascal VOC, oriented object detection on DOTA/STAR, and image classification on three common datasets. Exciting results illustrate that Mona surpasses full fine-tuning on all these tasks by tuning less than 5% params of the backbone, and is the only delta-tuning method outperforming full fine-tuning on all tasks. For example, Mona achieves 1% performance gain on the COCO compared to full fine-tuning. Comprehensive results suggest that Mona-tuning is more suitable for retaining and utilizing the capabilities of pre-trained models than full fine-tuning. The code is publicly available on https://github.com/Leiyi-Hu/mona.

****

Precise Event Spotting in Sports Videos: Solving Long-Range Dependency and Class Imbalance

Sanchayan Santra, Vishal Chudasama, Pankaj Wasnik, Vineeth N Balasubramanian; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3163-3172

Precise Event Spotting (PES) aims to identify events and their class from long, untrimmed videos, particularly in sports. The main objective of PES is to detect the event at the exact moment it occurs. Existing methods mainly rely on features from a large pre-trained network, which may not be ideal for the task. Furthermore, these methods overlook the issue of imbalanced event class distribution present in the data, negatively impacting performance in challenging scenarios. This paper demonstrates that an appropriately designed network, trained end-to-end, can outperform state-of-the-art (SOTA) methods. Particularly, we propose a network with a convolutional spatial-temporal feature extractor enhanced with our

proposed Adaptive Spatio-Temporal Refinement Module (ASTRM) and a long-range tem
poral module. The ASTRM enhances the features with spatio-temporal information.
Meanwhile, the long-range temporal module helps extract global context from the
data by modeling long-range dependencies. To address the class imbalance issue,
we introduce the Soft Instance Contrastive (Soft-IC) loss that promotes feature
compactness and class separation. Extensive experiments show that the proposed m
ethod is efficient and outperforms the SOTA methods, specifically in more challe
nging settings.
****

Real-IAD D3: A Real-World 2D/Pseudo-3D/3D Dataset for Industrial Anomaly Detecti
on

Wenbing Zhu, Lidong Wang, Ziqing Zhou, Chengjie Wang, Yurui Pan, Ruoyi Zhang, Zh
uhao Chen, Linjie Cheng, Bin-Bin Gao, Jiangning Zhang, Zhenye Gan, Yuxie Wang, Y
ulong Chen, Shuguang Qian, Mingmin Chi, Bo Peng, Lizhuang Ma; Proceedings of the
 Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15214-1522
3

The increasing complexity of industrial anomaly detection (IAD) has positioned m
ultimodal detection methods as a focal area of machine vision research. However,
 dedicated multimodal datasets specifically tailored for IAD remain limited. Pio
neering datasets like MVTec 3D have laid essential groundwork in multimodal IAD
by incorporating RGB+3D data, but still face challenges in bridging the gap with
 real industrial environments due to limitations in scale and resolution. To add
ress these challenges, we introduce Real-IAD D3, a high-precision multimodal dat
aset that uniquely incorporates an additional pseudo-3D modality generated throu
gh photometric stereo, alongside high-resolution RGB images and micrometer-level
 3D point clouds. Real-IAD D3 comprises industrial components with smaller dimen
sions and finer defects than existing datasets, offering diverse anomalies acros
s modalities and presenting a more challenging benchmark for multimodal IAD rese
arch. With 20 product categories, the dataset offers significantly greater scale
 and diversity compared to current alternatives. Additionally, we introduce an e
ffective approach that integrates RGB, point cloud, and pseudo-3D depth informat
ion to leverage the complementary strengths of each modality, enhancing detectio
n performance. Our experiments highlight the importance of these modalities in b
oosting detection robustness and overall IAD performance. The Real-IAD D3 datase
t will be publicly available to advance research and innovation in multimodal IA
D.The dataset and code are publicly accessible for research purposes at https://
realiad4ad.github.io/Real-IAD_D3.
****

Steady Progress Beats Stagnation: Mutual Aid of Foundation and Conventional Mode
ls in Mixed Domain Semi-Supervised Medical Image Segmentation

Qinghe Ma, Jian Zhang, Zekun Li, Lei Qi, Qian Yu, Yinghuan Shi; Proceedings of t
he Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5175-518
5

Large pretrained visual foundation models exhibit impressive general capabilitie
s. However, the extensive prior knowledge inherent in these models can sometimes
 be a double-edged sword when adapting them to downstream tasks in specific doma
ins.In the context of semi-supervised medical image segmentation with domain shi
ft, foundation models like MedSAM tend to make overconfident predictions, some o
f which are incorrect. The error accumulation hinders the effective utilization
of unlabeled data and limits further improvements.In this paper, we introduce a
Synergistic training framework for Foundation and Conventional models (SynFoC) t
o address the issue. We observe that a conventional model trained from scratch h
as the ability to correct the high-confidence mispredictions of the foundation m
odel, while the foundation model can supervise it with high-quality pseudo-label
s in the early training stages. Furthermore, to enhance the collaborative traini
ng effectiveness of both models and promote reliable convergence towards optimiz
ation, the consensus-divergence consistency regularization is proposed. We demon
strate the superiority of our method across four public multi-domain datasets. I
n particular, our method improves the Dice score by 10.31% on the Prostate datas
et. Our code is available in the supplementary material.

****
ATP: Adaptive Threshold Pruning for Efficient Data Encoding in Quantum Neural Networks
Mohamed Afane, Gabrielle Ebbrecht, Ying Wang, Juntao Chen, Junaid Farooq; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20427-20436
Quantum Neural Networks (QNNs) offer promising capabilities for complex data tasks, but are often constrained by limited qubit resources and high entanglement, which can hinder scalability and efficiency. In this paper, we introduce Adaptive Threshold Pruning (ATP), an encoding method that reduces entanglement and optimizes data complexity for efficient computations in QNNs. ATP dynamically prunes non-essential features in the data based on adaptive thresholds, effectively reducing quantum circuit requirements while preserving high performance. Extensive experiments across multiple datasets demonstrate that ATP reduces entanglement entropy and improves adversarial robustness when combined with adversarial training methods like FGSM. Our results highlight ATP's ability to balance computational efficiency and model resilience, achieving significant performance improvements with fewer resources, which will help make QNNs more feasible in practical, resource-constrained settings.
****
Color Alignment in Diffusion
Ka Chun Shum, Binh-Son Hua, Duc Thanh Nguyen, Sai-Kit Yeung; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28446-28455
Diffusion models have shown great promise in synthesizing visually appealing images. However, it remains challenging to condition the synthesis at a fine-grained level, for instance, synthesizing image pixels following some generic color pattern. Existing image synthesis methods often produce contents that fall outside the desired pixel conditions. To address this, we introduce a novel color alignment algorithm that confines the generative process in diffusion models within a given color pattern. Specifically, we project diffusion terms, either imagery samples or latent representations, into a conditional color space to align with the input color distribution. This strategy simplifies the prediction in diffusion models within a color manifold while still allowing plausible structures in generated contents, thus enabling the generation of diverse contents that comply with the target color pattern. Experimental results demonstrate our state-of-the-art performance in conditioning and controlling of color pixels, while maintaining on-par generation quality and diversity in comparison with regular diffusion models.
****
LLAVIDAL: A Large LAnguage VIsion Model for Daily Activities of Living
Dominick Reilly, Rajatsubhra Chakraborty, Arkaprava Sinha, Manish Kumar Govind, Pu Wang, Francois Bremond, Le Xue, Srijan Das; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24297-24308
Current Large Language Vision Models (LLVMs) trained on web videos perform well in general video understanding but struggle with fine-grained details, complex human-object interactions (HOI), and view-invariant representation learning essential for Activities of Daily Living (ADL). This limitation stems from a lack of specialized ADL video instruction-tuning datasets and insufficient modality integration to capture discriminative action representations. To address this, we propose a semi-automated framework for curating ADL datasets, creating ADL-X, a multi-view, multi-modal RGBS instruction-tuning dataset. Additionally, we introduce LLAVIDAL, an LLVM integrating videos, 3D skeletons, and HOIs to model ADL's complex spatiotemporal relationships. For training LLAVIDAL, a simple joint alignment of all modalities yields suboptimal results; thus, we propose a Multimodal Progressive (MMPro) training strategy, incorporating modalities in stages following a curriculum. We also establish ADL MCQ and video description benchmarks to assess LLVM performance in ADL tasks. Trained on ADL-X, LLAVIDAL achieves state-of-the-art performance across ADL benchmarks. Code and data will be made publicly available at https://adl-x.github.io/.
****

Language-Guided Salient Object Ranking
Fang Liu, Yuhao Liu, Ke Xu, Shuquan Ye, Gerhard Petrus Hancke, Rynson W. H. Lau; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29803-29813
Salient Object Ranking (SOR) aims to study human attention shifts across different objects in the scene. It is a challenging task, as it requires comprehension of the relations among the salient objects in the scene. However, existing works often overlook such relations or model them implicitly. In this work, we observe that when Large Vision-Language Models (LVLMs) describe a scene, they usually focus on the most salient object first, and then discuss the relations as they move on to the next (less salient) one. Based on this observation, we propose a novel Language-Guided Salient Object Ranking approach (named LG-SOR), which utilizes the internal knowledge within the LVLM-generated language descriptions, i.e., semantic relation cues and the implicit entity order cues, to facilitate saliency ranking. Specifically, we first propose a novel Text-Guided Visual Modulation (TGVM) module to incorporate semantic information in the description for saliency ranking. TGVM controls the flow of linguistic information to the visual features, suppresses noisy background image features, and enables propagation of useful textual features. We then propose a novel Text-Aware Visual Reasoning (TAVR) module to enhance model reasoning in object ranking, by explicitly learning a multimodal graph based on the entity and relation cues derived from the description. Extensive experiments demonstrate superior performances of our model on two SOR benchmarks.
****

Decoupled Motion Expression Video Segmentation
Hao Fang, Runmin Cong, Xiankai Lu, Xiaofei Zhou, Sam Kwong, Wei Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13821-13831
Motion expression video segmentation aims to segment objects based on input motion descriptions. Compared with traditional referring video object segmentation, it focuses on motion and multi-object expressions and is more challenging. Previous works achieved it by simply injecting text information into the video instance segmentation (VIS) model. However, this requires retraining the entire model and optimization is difficult. In this work, we propose DMVS, a simple framework constructed on the existing query-based VIS model, emphasizing decoupling the task into video instance segmentation and motion expression understanding. Firstly, we use a frozen video instance segmenter to extract object-specific contexts and convert them into frame-level and video-level queries. Secondly, we interact two levels of queries with static and motion cues, respectively, to further encode visually enhanced motion expressions. Furthermore, we propose a novel query initialization strategy that uses video queries guided by classification priors to initialize motion queries, greatly reducing the difficulty of optimization. Without bells and whistles, DMVS achieves state-of-the-art performance on the MeViS dataset at a lower training cost. Extensive experiments verify the effectiveness and efficiency of our framework.
****

K-LoRA: Unlocking Training-Free Fusion of Any Subject and Style LoRAs
Ziheng Ouyang, Zhen Li, Qibin Hou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13041-13050
Recent studies have explored combining different LoRAs to jointly generate learned style and content. However, existing methods either fail to effectively preserve both the original subject and style simultaneously or require additional training. In this paper, we argue that the intrinsic properties of LoRA can effectively guide diffusion models in merging learned subject and style. Building on this insight, we propose K-LoRA, a simple yet effective training-free LoRA fusion approach. In each attention layer, K-LoRA compares the Top-K elements in each LoRA to be fused, determining which LoRA to select for optimal fusion. This selection mechanism ensures that the most representative features of both subject and style are retained during the fusion process, effectively balancing their contributions. Experimental results demonstrate that the proposed method effectively i

ntegrates the subject and style information learned by the original LoRAs, outperforming state-of-the-art training-based approaches in both qualitative and quantitative results.
****

Towards More General Video-based Deepfake Detection through Facial Component Guided Adaptation for Foundation Model
Yue-Hua Han, Tai-Ming Huang, Kai-Lung Hua, Jun-Cheng Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22995-23005
The current deep generative models have enabled the creation of synthetic facial images with remarkable photorealism, raising significant societal concerns over their potential misuse. Despite rapid advancements in the field of deepfake detection, developing an efficient and effective approach for the generalized deepfake detection of unseen forgery samples remains challenging. To address this challenge, we leverage the rich semantic priors of foundation models and propose a novel side-network-based decoder that extracts spatial and temporal cues using the CLIP image encoder for generalized video-based Deepfake detection. Additionally, we introduce Facial Component Guidance (FCG) to enhance spatial learning generalizability by encouraging the model to focus on key facial regions. By leveraging the generic features of a vision-language foundation model, our approach demonstrates promising generalizability on challenging Deepfake datasets while also exhibiting superiority in training data efficiency, parameter efficiency, and model robustness. The source code is available at: https://github.com/aiiu-lab/DFD-FCG.
****

WF-VAE: Enhancing Video VAE by Wavelet-Driven Energy Flow for Latent Video Diffusion Model
Zongjian Li, Bin Lin, Yang Ye, Liuhan Chen, Xinhua Cheng, Shenghai Yuan, Li Yuan ; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17778-17788
Video Variational Autoencoder (VAE) encodes videos into a low-dimensional latent space, becoming a key component of most Latent Video Diffusion Models (LVDMs) to reduce model training costs. However, as the resolution and duration of generated videos increase, the encoding cost of Video VAEs becomes a limiting bottleneck in training LVDMs. Moreover, the block-wise inference method adopted by most LVDMs can lead to discontinuities of latent space when processing long-duration videos. The key to addressing the computational bottleneck lies in decomposing videos into distinct components and efficiently encoding the critical information. Wavelet transform can decompose videos into multiple frequency-domain components and improve the efficiency significantly, we thus propose Wavelet Flow VAE (WF-VAE), an autoencoder that leverages multi-level wavelet transform to facilitate low-frequency energy flow into latent representation. Furthermore, we introduce a method called Causal Cache, which maintains the integrity of latent space during block-wise inference. Compared to state-of-the-art video VAEs, WF-VAE demonstrates superior performance in both PSNR and LPIPS metrics, achieving 2x higher throughput and 4x lower memory consumption while maintaining competitive reconstruction quality. Our code and models are available at https://github.com/PKU-YuanGroup/WF-VAE.
****

SP3D: Boosting Sparsely-Supervised 3D Object Detection via Accurate Cross-Modal Semantic Prompts
Shijia Zhao, Qiming Xia, Xusheng Guo, Pufan Zou, Maoji Zheng, Hai Wu, Chenglu Wen, Cheng Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29374-29384
Recently, sparsely-supervised 3D object detection has gained great attention, achieving performance close to fully-supervised 3D detectors while requiring only a few annotated instances. Nevertheless, these methods suffer challenges when accurate labels are extremely absent. In this paper, we propose a boosting strategy, termed SP3D, explicitly utilizing the cross-modal semantic prompts generated from Large Multimodal Models (LMMs) to boost the 3D detector with robust feature discrimination capability under sparse annotation settings. Specifically, we fi

rst develop a Confident Points Semantic Transfer (CPST) module that generates ac
curate cross-modal semantic prompts through boundary-constrained center cluster
selection. Based on these accurate semantic prompts, which we treat as seed poin
ts, we introduce a Dynamic Cluster Pseudo-label Generation (DCPG) module to yiel
d pseudo-supervision signals from the geometry shape of multi-scale neighbor poi
nts. Additionally, we design a Distribution Shape score (DS score) that chooses
high-quality supervision signals for the initial training of the 3D detector. Ex
periments on the KITTI dataset and Waymo Open Dataset (WOD) have validated that
SP3D can enhance the performance of sparsely supervised detectors by a large mar
gin under meager labeling conditions. Moreover, we verified SP3D in the zero-sho
t setting, where its performance exceeded that of the state-of-the-art methods.
The code is available at https://github.com/xmuqimingxia/SP3D.
****

ARKit LabelMaker: A New Scale for Indoor 3D Scene Understanding
Guangda Ji, Silvan Weder, Francis Engelmann, Marc Pollefeys, Hermann Blum; Proce
edings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, p
p. 4398-4407
Neural network performance scales with both model size and data volume, as shown
 in both language and image processing. This requires scaling-friendly architect
ures and large datasets. While transformers have been adapted for 3D vision, a `
GPT-moment' remains elusive due to limited training data. We introduce ARKit Lab
elMaker, a large-scale real-world 3D dataset with dense semantic annotation that
 is more than three times larger than prior largest dataset. Specifically, we ex
tend ARKitScenes with automatically generated dense 3D labels using an extended
LabelMaker pipeline, tailored for large-scale pre-training. Training on our data
set improves accuracy across architectures, achieving state-of-the-art 3D semant
ic segmentation scores on ScanNet and ScanNet200, with notable gains on tail cla
sses. Our code is available at https://labelmaker.org and our dataset at https:/
/huggingface.co/datasets/labelmaker/arkit_labelmaker.
****

VoCo-LLaMA: Towards Vision Compression with Large Language Models
Xubing Ye, Yukang Gan, Xiaoke Huang, Yixiao Ge, Yansong Tang; Proceedings of the
 Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29836-2984
6
Vision-Language Models (VLMs) have achieved remarkable success in various multi-
modal tasks, but they are often bottlenecked by the limited context window and h
igh computational cost of processing high-resolution image inputs and videos. Vi
sion compression can alleviate this problem by reducing the vision token count.
Previous approaches compress vision tokens with external modules and force LLMs
to understand the compressed ones, leading to visual information loss. However,
the LLMs' understanding paradigm of vision tokens is not fully utilised in the c
ompression learning process. We propose VoCo-LLaMA, the first approach to compre
ss vision tokens using LLMs. By introducing Vision Compression tokens during the
 vision instruction tuning phase and leveraging attention distillation, our meth
od distill how LLMs comprehend vision tokens into their processing of VoCo token
s. VoCo-LLaMA facilitates effective vision compression and improves the computat
ional efficiency during the inference stage. Specifically, our method can achiev
e a 576x compression rate while maintaining 83.7% performance. Furthermore, thro
ugh continuous training using time-series compressed token sequences of video fr
ames, VoCo-LLaMA demonstrates the ability to understand temporal correlations, o
utperforming previous methods on popular video question-answering benchmarks. Ou
r approach presents a promising way to unlock the full potential of VLMs' contex
tual window, enabling more scalable multi-modal applications. The project page c
an be accessed via https://yxxxb.github.io/VoCo-LLaMA-page.
****

StreamingT2V: Consistent, Dynamic, and Extendable Long Video Generation from Tex
t
Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram
Tadevosyan, Zhangyang Wang, Shant Navasardyan, Humphrey Shi; Proceedings of the
Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2568-2577

Text-to-video diffusion models enable the generation of high-quality videos that follow text instructions, simplifying the process of producing diverse and individual content. Current methods excel in generating short videos (up to 16s), but produce hard-cuts when naively extended to long video synthesis. To overcome these limitations, we present StreamingT2V, an autoregressive method that generates long videos of up to 2 minutes or longer with seamless transitions. The key components are: (i) a short-term memory block called conditional attention module (CAM), which conditions the current generation on the features extracted from the preceding chunk via an attentional mechanism, leading to consistent chunk transitions, (ii) a long-term memory block called appearance preservation module (APM), which extracts high-level scene and object features from the first video chunk to prevent the model from forgetting the initial scene, and (iii) a randomized blending approach that allows for the autoregressive application of a video enhancer on videos of indefinite length, ensuring consistency across chunks. Experiments show that StreamingT2V produces more motion, while competing methods suffer from video stagnation when applied naively in an autoregressive fashion. Thus, we propose with StreamingT2V a high-quality seamless text-to-long video generator, surpassing competitors in both consistency and motion.
****

Focal Split: Untethered Snapshot Depth from Differential Defocus
Junjie Luo, John Mamish, Alan Fu, Thomas Concannon, Josiah Hester, Emma Alexander, Qi Guo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26965-26974
We introduce Focal Split, a handheld, snapshot depth camera with fully onboard power and computing based on depth-from-differential-defocus (DfDD). Focal Split is passive, avoiding power consumption of light sources. Its achromatic optical system simultaneously forms two differentially defocused images of the scene, which can be independently captured using two photosensors in a snapshot. The data processing is based on the DfDD theory, which efficiently computes a depth and a confidence value for each pixel with only 500 floating point operations (FLOPs) per pixel from the camera measurements. We demonstrate a Focal Split prototype, which comprises a handheld custom camera system connected to a Raspberry Pi 5 for real-time data processing. The system consumes 4.9 W and is powered on a 5 V, 10,000 mAh battery. The prototype can measure objects with distances from 0.4 m to 1.2 m, outputting 480 x 360 sparse depth maps at 2.1 frames per second (FPS) using unoptimized Python scripts. Focal Split is DIY friendly. A comprehensive guide to building your own Focal Split depth camera, code, and additional data can be found at https://focal-split.qiguo.org.
****

AFL: A Single-Round Analytic Approach for Federated Learning with Pre-trained Models
Run He, Kai Tong, Di Fang, Han Sun, Ziqian Zeng, Haoran Li, Tianyi Chen, Huiping Zhuang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4988-4998
In this paper, we introduce analytic federated learning (AFL), a new training paradigm that brings analytical (i.e., closed-form) solutions to the federated learning (FL) with pre-trained models. Our AFL draws inspiration from analytic learning---a gradient-free technique that trains neural networks with analytical solutions in one epoch. In the local client training stage, the AFL facilitates a one-epoch training, eliminating the necessity for multi-epoch updates. In the aggregation stage, we derive an absolute aggregation (AA) law. This AA law allows a single-round aggregation, reducing heavy communication overhead and achieving fast convergence by removing the need for multiple aggregation rounds. More importantly, the AFL exhibits a property that invariance to data partitioning, meaning that regardless of how the full dataset is distributed among clients, the aggregated result remains identical. This could spawn various potentials, such as data heterogeneity invariance and client-number invariance. We conduct experiments across various FL settings including extremely non-IID ones, and scenarios with a large number of clients (e.g., \ge 1000). In all these settings, our AFL constantly performs competitively while existing FL techniques encounter various ob

stacles.
****
XLRS-Bench: Could Your Multimodal LLMs Understand Extremely Large Ultra-High-Resolution Remote Sensing Imagery?
Fengxiang Wang, Hongzhen Wang, Zonghao Guo, Di Wang, Yulin Wang, Mingshuo Chen, Qiang Ma, Long Lan, Wenjing Yang, Jing Zhang, Zhiyuan Liu, Maosong Sun; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14325-14336
The astonishing breakthrough of multimodal large language models (MLLMs) has necessitated new benchmarks to quantitatively assess their capabilities, reveal their limitations, and indicate future research directions. However, this is challenging in the context of remote sensing (RS), since the imagery features ultra-high resolution that incorporates extremely complex semantic relationships. Existing benchmarks usually adopt notably smaller image sizes than real-world RS scenarios, suffer from limited annotation quality, and consider insufficient dimensions of evaluation. To address these issues, we present XLRS-Bench: a comprehensive benchmark for evaluating the perception and reasoning capabilities of MLLMs in ultra-high-resolution RS scenarios. XLRS-Bench boasts the largest average image size (8500x8500) observed thus far, with all evaluation samples meticulously annotated manually, assisted by a novel semi-automatic captioner on ultra-high-resolution RS images. On top of the XLRS-Bench, 16 sub-tasks are defined to evaluate MLLMs' 10 kinds of perceptual capabilities and 6 kinds of reasoning capabilities, with a primary emphasis on advanced cognitive processes that facilitate real-world decision-making and the capture of spatiotemporal changes. The results of both general and RS-focused MLLMs on XLRS-Bench indicate that further efforts are needed for real-world RS applications. We will open source XLRS-Bench to support further research of developing more powerful MLLMs for RS.
****
BOLT: Boost Large Vision-Language Model Without Training for Long-form Video Understanding
Shuming Liu, Chen Zhao, Tianqi Xu, Bernard Ghanem; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3318-3327
Large video-language models (VLMs) have demonstrated promising progress in various video understanding tasks. However, their effectiveness in long-form video analysis is constrained by limited context windows. Traditional approaches, such as uniform frame sampling, often inevitably allocate resources to irrelevant content, diminishing their effectiveness in real-world scenarios. In this paper, we introduce BOLT, a method to boost large VLMs without additional training through a comprehensive study of frame selection strategies. First, to enable a more realistic evaluation of VLMs in long-form video understanding, we propose a multi-source retrieval evaluation setting. Our findings reveal that uniform sampling performs poorly in noisy contexts, underscoring the importance of selecting the right frames. Second, we explore several frame selection strategies based on query-frame similarity and analyze their effectiveness at inference time. Our results show that inverse transform sampling yields the most significant performance improvement, increasing accuracy on the Video-MME benchmark from 53.8% to 56.1% and MLVU benchmark from 58.9% to 63.4%.
****
Efficient Data Driven Mixture-of-Expert Extraction from Trained Networks
Uranik Berisha, Jens Mehnert, Alexandru Paul Condurache; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20082-20091
Vision Transformers (ViTs) have emerged as the state-of-the-art models in various Computer Vision (CV) tasks, but their high computational and resource demands pose significant challenges. While Mixture of Experts (MoE) can make these models more efficient, they often require costly retraining or even training from scratch. Recent developments aim to reduce these computational costs by leveraging pretrained networks. These have been shown to produce sparse activation patterns in the Multi-Layer Perceptrons (MLPs) of the encoder blocks, allowing for conditional activation of only relevant subnetworks for each sample. Building on this idea, we propose a new method to construct MoE variants from pretrained models.

Our approach extracts expert subnetworks from the model's MLP layers post-training in two phases. First, we cluster output activations to identify distinct activation patterns. In the second phase, we use these clusters to extract the corresponding subnetworks responsible for producing them. On ImageNet-1k recognition tasks, we demonstrate that these extracted experts can perform surprisingly well out of the box and require only minimal fine-tuning to regain 98% of the original performance, all while reducing FLOPs and model size, by up to 36% and 32% respectively.

****

Reference-Based 3D-Aware Image Editing with Triplanes

Bahri Batuhan Bilecen, Yigit Yalin, Ning Yu, Aysegul Dundar; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5904-5915

Generative Adversarial Networks (GANs) have emerged as powerful tools for high-quality image generation and real image editing by manipulating their latent spaces. Recent advancements in GANs include 3D-aware models such as EG3D, which feature efficient triplane-based architectures capable of reconstructing 3D geometry from single images. However, limited attention has been given to providing an integrated framework for 3D-aware, high-quality, reference-based image editing. This study addresses this gap by exploring and demonstrating the effectiveness of the triplane space for advanced reference-based edits. Our novel approach integrates encoding, automatic localization, spatial disentanglement of triplane features, and fusion learning to achieve the desired edits. We demonstrate how our approach excels across diverse domains, including human faces, 360-degree heads, animal faces, partially stylized edits like cartoon faces, full-body clothing edits, and edits on class-agnostic samples. Our method shows state-of-the-art performance over relevant latent direction, text, and image-guided 2D and 3D-aware diffusion and GAN methods, both qualitatively and quantitatively.

****

StyleSSP: Sampling StartPoint Enhancement for Training-free Diffusion-based Method for Style Transfer

Ruojun Xu, Weijie Xi, XiaoDi Wang, Yongbo Mao, Zach Cheng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18260-18269

Training-free diffusion-based methods have achieved remarkable success in style transfer, eliminating the need for extensive training or fine-tuning. However, due to the lack of targeted training for style information extraction and constraints on the content image layout, training-free methods often suffer from layout changes of original content and content leakage from style images. Through a series of experiments, we discovered that an effective startpoint in the sampling stage significantly enhances the style transfer process. Based on this discovery, we propose StyleSSP, which focuses on obtaining a better startpoint to address layout changes of original content and content leakage from style image. StyleSSP comprises two key components: (1) Frequency Manipulation: To improve content preservation, we reduce the low-frequency components of the DDIM latent, allowing the sampling stage to pay more attention to the layout of content images; and (2) Negative Guidance via Inversion: To mitigate the content leakage from style image, we employ negative guidance in the inversion stage to ensure that the startpoint of the sampling stage is distanced from the content of style image. Experiments show that StyleSSP surpasses previous training-free style transfer baselines, particularly in preserving original content and minimizing the content leakage from style image.

****

PURA: Parameter Update-Recovery Test-Time Adaption for RGB-T Tracking

Zekai Shao, Yufan Hu, Bin Fan, Hongmin Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22089-22098

Maintaining stable tracking of objects in domain shift scenarios is crucial for RGB-T tracking, prompting us to explore the use of unlabeled test sample information for effective online model adaptation. However, current Test-Time Adaptation (TTA) methods in RGB-T tracking dramatically change the model's internal parameters during long-term adaptation. At the same time, the gradient computations involved in the optimization process impose a significant computational burden. T

o address these challenges, we propose a Parameter Update-Recovery Adaptation (PURA) framework based on parameter decomposition. Firstly, our fast parameter update strategy adjusts model parameters using statistical information from test samples without requiring gradient calculations, ensuring consistency between the model and test data distribution. Secondly, our parameter decomposition recovery employs orthogonal decomposition to identify the principal update direction and recover parameters in this direction, aiding in the retention of critical knowledge. Finally, we leverage the information obtained from decomposition to provide feedback on the momentum during the update phase, ensuring a stable updating process. Experimental results demonstrate that PURA outperforms current state-of-the-art methods across multiple datasets, validating its effectiveness. The project page is at https://melantech.github.io/PURA.

****

One is Plenty: A Polymorphic Feature Interpreter for Immutable Heterogeneous Collaborative Perception

Yuchen Xia, Quan Yuan, Guiyang Luo, Xiaoyuan Fu, Yang Li, Xuanhan Zhu, Tianyou Luo, Siheng Chen, Jinglin Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1592-1601

Collaborative perception in autonomous driving significantly enhances the perception capabilities of individual agents. Immutable heterogeneity in collaborative perception, where agents have different and fixed perception networks, presents a major challenge due to the semantic gap in their exchanged intermediate features without modifying the perception networks. Most existing methods bridge the semantic gap through interpreters. However, they either require training a new interpreter for each new agent type, limiting extensibility, or rely on a two-stage interpretation via an intermediate standardized semantic space, causing cumulative semantic loss. To achieve both extensibility in immutable heterogeneous scenarios and low-loss feature interpretation, we propose PolyInter, a polymorphic feature interpreter. It contains an extension point through which emerging new agents can seamlessly integrate by overriding only their specific prompts, which are learnable parameters intended to guide the interpretation, while reusing PolyInter's remaining parameters. By leveraging polymorphism, our design ensures that single interpreter is sufficient to accommodate diverse agents and interpret their features into the ego agent's semantic space. Experiments conducted on OPV2V dataset demonstrate that PolyInter improves collaborative perception precision by up to 11.1% compared to SOTA interpreters, while comparable results can be achieved by training only 1.4% of PolyInter's parameters when adapting to new agents.

****

Towards All-in-One Medical Image Re-Identification

Yuan Tian, Kaiyuan Ji, Rongzhao Zhang, Yankai Jiang, Chunyi Li, Xiaosong Wang, Guangtao Zhai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30774-30786

Medical image re-identification (MedReID) is under-explored so far, despite its critical applications in personalized healthcare and privacy protection.In this paper, we introduce a thorough benchmark and a unified model for this problem.First, to handle various medical modalities, we propose a novel Continuous Modality-based Parameter Adapter (ComPA). ComPA condenses medical content into a continuous modality representation and dynamically adjusts the modality-agnostic model with modality-specific parameters at runtime. This allows a single model to adaptively learn and process diverse modality data.Furthermore, we integrate medical priors into our model by aligning it with a bag of pre-trained medical foundation models, in terms of the differential features.Compared to single-image feature, modeling the inter-image difference better fits the re-identification problem, which involves discriminating multiple images.We evaluate the proposed model against 25 foundation models and 8 large multi-modal language models across 11 image datasets, demonstrating consistently superior performance.Additionally, we deploy the proposed MedReID technique to two real-world applications, i.e., history-augmented personalized diagnosis and medical privacy protection.

****

SegAgent: Exploring Pixel Understanding Capabilities in MLLMs by Imitating Human Annotator Trajectories
Muzhi Zhu, Yuzhuo Tian, Hao Chen, Chunluan Zhou, Qingpei Guo, Yang Liu, Ming Yang, Chunhua Shen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3686-3696
While MLLMs have demonstrated adequate image understanding capabilities, they still struggle with pixel-level comprehension, limiting their practical applications. Current evaluation tasks like VQA and visual grounding remain too coarse to assess fine-grained pixel comprehension accurately. Though segmentation is foundational for pixel-level understanding, existing methods often require MLLMs to generate implicit tokens, decoded through external pixel decoders. This approach disrupts the MLLM's text output space, potentially compromising language capabilities and reducing flexibility and extensibility, while failing to reflect the model's intrinsic pixel-level understanding.Thus, We introduce the Human-Like Mask Annotation Task (HLMAT), a new paradigm where MLLMs mimic human annotators using interactive segmentation tools. Modeling segmentation as a multi-step Markov Decision Process, HLMAT enables MLLMs to iteratively generate text-based click points, achieving high-quality masks without architectural changes or implicit tokens. Through this setup, we develop SegAgent, a model fine-tuned on human-like annotation trajectories, which achieves performance comparable to SOTA methods and supports additional tasks like mask refinement and annotation filtering.HLMAT provides a protocol for assessing fine-grained pixel understanding in MLLMs and introduces a vision-centric, multi-step decision-making task that facilitates exploration of MLLMs' visual reasoning abilities. Our adaptations of policy improvement method StaR and PRM guided tree search further enhance model robustness in complex segmentation tasks, laying a foundation for future advancements in fine-grained visual perception and multi-step decision-making for MLLMs.
****

Motions as Queries: One-Stage Multi-Person Holistic Human Motion Capture
Kenkun Liu, Yurong Fu, Weihao Yuan, Jing Lin, Peihao Li, Xiaodong Gu, Lingteng Qiu, Haoqian Wang, Zilong Dong, Xiaoguang Han; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17529-17539
Existing methods for capturing multi-person holistic human motions from a monocular video usually involve integrating the detector, the tracker, and the human pose & shape estimator into a cascaded system. Differently, we develop a one-stage multi-person holistic human motion capture system, which 1) employs only one network, enabling significant benefits from the end-to-end training on a large-scale dataset; 2) enables performance improving of the tracking module during training, avoiding being limited by a pre-trained tracker; 3) captures the motions of all individuals within a single shot, rather than tracking and estimating each person sequentially. In this system, each query within a temporal cross-attention module is responsible for the long motion of a specific individual, implicitly aggregating individual-specific information throughout the entire video. To further boost the proposed system from end-to-end training, we also construct a synthetic human video dataset, with multi-person and whole-body annotations. Extensive experiments across different datasets demonstrate both the efficacy and the efficiency of both the proposed method and the dataset. The code of our method will be made publicly available.
****

SceneCrafter: Controllable Multi-View Driving Scene Editing
Zehao Zhu, Yuliang Zou, Chiyu Max Jiang, Bo Sun, Vincent Casser, Xiukun Huang, Jiahao Wang, Zhenpei Yang, Ruiqi Gao, Leonidas Guibas, Mingxing Tan, Dragomir Anguelov; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6812-6822
Simulation is crucial for developing and evaluating autonomous vehicle (AV) systems. Recent literature builds on a new generation of generative models to synthesize highly realistic images for full-stack simulation. However, purely synthetically generated scenes are not grounded in reality and have difficulty in inspiring confidence in the relevance of its outcomes. Editing models, on the other hand, leverage source scenes from real driving logs, and enable the simulation of

different traffic layouts, behaviors, and operating conditions such as weather a
nd time of day. While image editing is an established topic in computer vision,
it presents fresh sets of challenges in driving simulation: (1) the need for cro
ss-camera 3D consistency, (2) learning "empty street" priors from driving data w
ith foreground occlusions, and (3) obtaining paired image tuples of varied editi
ng conditions while preserving consistent layout and geometry. To address these
challenges, we propose SceneCrafter, a versatile editor for realistic 3D-consist
ent manipulation of driving scenes captured from multiple cameras. We build on r
ecent advancements in multi-view diffusion models, using a fully controllable fr
amework that scales seamlessly to multi-modality conditions like weather, time o
f day, agent boxes and high-definition maps. To generate paired data for supervi
sing the editing model, we propose a novel framework on top of Prompt-to-Prompt
to generate geometrically consistent synthetic paired data with global edits. We
 also introduce an alpha-blending framework to synthesize data with local edits,
 leveraging a model trained on empty street priors through novel masked training
 and multi-view repaint paradigm. SceneCrafter demonstrates powerful editing cap
abilities and achieves state-of-the-art realism, controllability, 3D consistency
, and scene editing quality compared to existing baselines.
****

AMO Sampler: Enhancing Text Rendering with Overshooting
Xixi Hu, Keyang Xu, Bo Liu, Qiang Liu, Hongliang Fei; Proceedings of the Compute
r Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13157-13166
Achieving precise alignment between textual instructions and generated images in
 text-to-image generation is a significant challenge, particularly in rendering
written text within images. Open-source models like Stable Diffusion 3 (SD3), Fl
ux, and AuraFlow often struggle with accurate text depiction, resulting in missp
elled or inconsistent text. We introduce a training-free method with minimal com
putational overhead that significantly enhances text rendering quality. Specific
ally, we introduce an overshooting sampler for a pretrained RF model, by alterna
ting between over-simulating the learned ordinary differential equation (ODE) an
d reintroducing noise. Compared to the Euler sampler, the overshooting sampler e
ffectively introduces an extra Langevin dynamics term that can help correct the
compounding error from successive Euler steps and therefore improve the text ren
dering. However, when the overshooting strength is high, we observe over-smoothi
ng artifacts on the generated images. To address this issue, we adaptively contr
ol the strength of the overshooting for each image patch according to their atte
ntion score with the text content. We name the proposed sampler Attention Modula
ted Overshooting sampler (AMO). AMO demonstrates a 32.3% and 35.9% improvement i
n text rendering accuracy on SD3 and Flux without compromising overall image qua
lity or increasing inference cost.
****

ImViD: Immersive Volumetric Videos for Enhanced VR Engagement
Zhengxian Yang, Shi Pan, Shengqi Wang, Haoxiang Wang, Li Lin, Guanjun Li, Zhengq
i Wen, Borong Lin, Jianhua Tao, Tao Yu; Proceedings of the Computer Vision and P
attern Recognition Conference (CVPR), 2025, pp. 16554-16564
User engagement is greatly enhanced by fully immersive multimodal experiences th
at combine visual and auditory stimuli. Consequently, the next frontier in VR/AR
 technologies lies in immersive volumetric videos with complete scene capture, l
arge 6-DoF interactive space, Multi-modal feedback, and high resolution & frame-
rate contents. To stimulate the reconstruction of immersive volumetric videos, w
e introduce ImViD, a multi-view, multi-modal dataset featuring complete space-or
iented data capture and various indoor/outdoor scenarios. Our capture rig suppor
ts multi-view video-audio capture while on the move, a capability absent in exis
ting datasets, which significantly enhances the completeness, flexibility, and e
fficiency of data capture. The captured multi-view videos (with synchronized aud
ios) are in 5K resolution at 60FPS, lasting from 1-5 minutes, and include rich f
oreground-background elements, and complex dynamics. We benchmark existing metho
ds using our dataset and establish a base pipeline for constructing immersive vo
lumetric videos from multi-view audiovisual inputs for 6-DoF multimodal immersiv
e VR experiences. The benchmark and the reconstruction and interaction results d

emonstrate the effectiveness of our dataset and baseline method, which we believ e will stimulate future research on immersive volumetric video production.
****

Integral Fast Fourier Color Constancy
Wenjun Wei, Yanlin Qian, Huaian Chen, Junkang Dai, Yi Jin; Proceedings of the Co mputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26420-26429
Traditional auto white balance (AWB) algorithms typically assume a single global illuminant source, which leads to color distortions in multi-illuminant scenes. While recent neural network-based methods have shown excellent accuracy in such scenarios, their high parameter count and computational demands limit their pra cticality for real-time video applications. The Fast Fourier Color Constancy (FF CC) algorithm was proposed for single-illuminant-source scenes, predicting a glo bal illuminant source with high efficiency. However, it cannot be directly appli ed to multi-illuminant scenarios unless specifically modified. To address this, we propose Integral Fast Fourier Color Constancy (IFFCC), an extension of FFCC t ailored for multi-illuminant scenes. IFFCC leverages the proposed integral UV hi stogram to accelerate histogram computations across all possible regions in Cart esian space and parallelizes Fourier-based convolution operations, resulting in a spatially-smooth illumination map. This approach enables high-accuracy, real-t ime AWB in multi-illuminant scenes. Extensive experiments show that IFFCC achiev es accuracy that is on par with or surpasses that of pixel-level neural networks , while reducing the parameter count by over 400x and processing speed by 20 - 1 00x faster than network-based approaches.
****

I2VGuard: Safeguarding Images against Misuse in Diffusion-based Image-to-Video M odels
Dongnan Gui, Xun Guo, Wengang Zhou, Yan Lu; Proceedings of the Computer Vision a nd Pattern Recognition Conference (CVPR), 2025, pp. 12595-12604
Recent advances in image-to-video generation have enabled animation of still ima ges and offered pixel-level controllability. While these models hold great poten tial to transform single images into vivid and dynamic videos, they also carry r isks of misuse that could impact privacy, security, and copyright protection. Th is paper proposes a novel approach that applies imperceptible perturbations on i mages to degrade the quality of the generated videos, thereby protecting images from misuse in white-box image-to-video diffusion models. Specifically, we funct ion our approach as an adversarial attack, incorporating spatial, temporal, and diffusion attack modules. The spatial attack shifts image features from their o riginal distribution to a lower-quality target distribution, reducing visual fid elity. The temporal attack disrupts coherent motion by interfering with temporal attention maps that guide motion generation. To enhance the robustness of our a pproach across different models, we further propose a diffusion attack module le veraging contrastive loss. Our approach can be easily integrated with mainstream diffusion-based I2V models. Extensive experiments on SVD, CogVideoX, and Contro lNeXt demonstrate that our method significantly impairs generation quality in te rms of visual clarity and motion consistency, while introducing only minimal art ifacts to the images. To the best of our knowledge, we are the first to explore adversarial attacks on image-to-video generation for security purposes.
****

Saliuitl: Ensemble Salience Guided Recovery of Adversarial Patches against CNNs
Mauricio Byrd Victorica, György Dán, Henrik Sandberg; Proceedings of the Compute r Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20360-20369
Adversarial patches are capable of misleading computer vision systems based on c onvolutional neural networks. Existing recovery methods suffer of at least one o f three fundamental shortcomings: no information about the presence of patches i n the scene, inability to efficiently handle noncontiguous patch attacks, and a strong reliance on fixed saliency thresholds. We propose Saliuitl, a recovery m ethod independent of the number of patches and their shape, which unlike prior w orks, explicitly detects patch attacks before attempting recovery. In our approa ch, detection is based on the attributes of a binarized feature map ensemble, wh ich is generated by using an ensemble of saliency thresholds. If an attack is de

tected, Saliuitl recovers clean predictions locating patches guided by an ensemb le of binarized feature maps and inpainting them. We evaluate Saliuitl on widely used object detection and image classification benchmarks from the adversarial patch literature, and our results show that compared to recent state-of-the-art defenses, Saliuitl achieves a recovery rate up to 97.81 and 42.63 percentage po ints higher at the same rate of lost predictions for image classification and ob ject detection, respectively. By design, Saliuitl has low computational complexi ty and is robust to adaptive white-box attacks. Our code is available at https:/ /github.com/Saliuitl/Saliuitl/tree/main.
****

HeatFormer: A Neural Optimizer for Multiview Human Mesh Recovery
Yuto Matsubara, Ko Nishino; Proceedings of the Computer Vision and Pattern Recog nition Conference (CVPR), 2025, pp. 6415-6424
We introduce a novel method for human shape and pose recovery that can fully lev erage multiple static views. We target fixed-multiview people monitoring, includ ing elderly care and safety monitoring, in which cameras can be installed at the corners of a room or an open space but whose configuration may vary depending o n the environment. Our key idea is to formulate it as neural optimization. We ac hieve this with HeatFormer, a neural optimizer that iteratively refines the SMPL parameters given multiview images. HeatFormer realizes this SMPL parameter esti mation as heatmap generation and alignment with a novel transformer encoder and decoder. Our formulation makes HeatFormer fundamentally agnostic to the number o f cameras, their configuration, and calibration. We demonstrate the effectivenes s of HeatFormer including its accuracy, robustness to occlusion, and generalizab ility through an extensive set of experiments. We believe HeatFormer can serve a key role in passive human behavior modeling.
****

ResCLIP: Residual Attention for Training-free Dense Vision-language Inference
Yuhang Yang, Jinhong Deng, Wen Li, Lixin Duan; Proceedings of the Computer Visio n and Pattern Recognition Conference (CVPR), 2025, pp. 29968-29978
While vision-language models like CLIP have shown remarkable success in open-voc abulary tasks, their application is currently confined to image-level tasks, and they still struggle with dense predictions. Recent works often attribute such d eficiency in dense predictions to the self-attention layers in the final block, and have achieved commendable results by modifying the original query-key attent ion to self-correlation attention, (e.g., query-query and key-key attention). Ho wever, these methods overlook the cross-correlation attention (query-key) proper ties, which capture the rich spatial correspondence. In this paper, we reveal th at the cross-correlation of self-attention in non-final layers of CLIP also exhi bits localization properties. Therefore, we propose the Residual Cross-correlati on Self-attention (RCS) module, which leverages the cross-correlation self-atten tion from intermediate layers to remold the attention in the final block. The RC S module effectively reorganizes spatial information, unleashing the localizatio n potential within CLIP for dense vision-language inference. Furthermore, to enh ance the focus on regions of the same categories and local consistency, we propo se the Semantic Feedback Refinement (SFR) module, which utilizes semantic segmen tation maps to further adjust the attention scores. By integrating these two str ategies, our method, termed ResCLIP, can be easily incorporated into existing ap proaches as a plug-and-play module, significantly boosting their performance in dense vision-language inference. Extensive experiments across multiple standard benchmarks demonstrate that our method surpasses state-of-the-art training-free methods, validating the effectiveness of the proposed approach. Code is availabl e at https://github.com/yvhangyang/ResCLIP.
****

GPS as a Control Signal for Image Generation
Chao Feng, Ziyang Chen, Aleksander Holynski, Alexei A. Efros, Andrew Owens; Proc eedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2766-2778
We show that the GPS tags contained in photo metadata provide a useful control s ignal for image generation. We train GPS-to-image models and use them for tasks

that require a fine-grained understanding of how images vary within a city. In particular, we train a diffusion model to generate images conditioned on both GPS and text. The learned model generates images that capture the distinctive appearance of different neighborhoods, parks, and landmarks. We also extract 3D models from 2D GPS-to-image models through score distillation sampling, using GPS conditioning to constrain the appearance of the reconstruction from each viewpoint. Our evaluations suggest that our GPS-conditioned models successfully learn to generate images that vary based on location, and that GPS conditioning improves estimated 3D structure.
****

CPath-Omni: A Unified Multimodal Foundation Model for Patch and Whole Slide Image Analysis in Computational Pathology

Yuxuan Sun, Yixuan Si, Chenglu Zhu, Xuan Gong, Kai Zhang, Pingyi Chen, Ye Zhang, Zhongyi Shui, Tao Lin, Lin Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10360-10371

The emergence of large multimodal models (LMMs) has brought significant advancements to pathology. Previous research has primarily focused on separately training patch-level and whole-slide image (WSI)-level models, limiting the integration of learned knowledge across patches and WSIs and resulting in redundant models. In this work, we introduce CPath-Omni, the first 15B parameter LMM that unifies patch and WSI analysis, consolidating a variety of tasks at both levels, including classification, visual question answering, captioning, and visual referring prompting. Extensive experiments demonstrate that CPath-Omni achieves state-of-the-art (SOTA) performance across seven diverse tasks on 39 out of 42 datasets, outperforming or matching task-specific models trained for individual tasks. Additionally, we develop a specialized pathology CLIP-based visual processor for CPath-Omni, CPath-CLIP, which, for the first time, integrates different vision models and incorporates a large language model as a text encoder to build a more powerful CLIP model, which achieves SOTA performance on nine zero-shot and four few-shot datasets. Our findings highlight CPath-Omni's ability to unify diverse pathology tasks, demonstrating its potential to streamline and advance the field of foundation model in pathology. The code and model are available at https://github.com/PathFoundation/CPath-Omni.
****

OPTICAL: Leveraging Optimal Transport for Contribution Allocation in Dataset Distillation

Xiao Cui, Yulei Qin, Wengang Zhou, Hongsheng Li, Houqiang Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15245-15254

The demands for increasingly large-scale datasets pose substantial storage and computation challenges to building deep learning models. Dataset distillation methods, especially those via sample generation techniques, rise in response to condensing large original datasets into small synthetic ones while preserving critical information. Existing subset synthesis methods simply minimize the homogeneous distance where uniform contributions from all real instances are allocated to shaping each synthetic sample. We demonstrate that such equal allocation fails to consider the instance-level relationship between each real-synthetic pair and gives rise to insufficient modeling of geometric structural nuances between the distilled and original sets. In this paper, we propose a novel framework named OPTICAL to reformulate the homogeneous distance minimization into a bi-level optimization problem via matching-and-approximating. In the matching step, we leverage optimal transport matrix to dynamically allocate contributions from real instances. Subsequently, we polish the generated samples in accordance with the established allocation scheme for approximating the real ones. Such a strategy better measures intricate geometric characteristics and handles intra-class variations for high fidelity of data distillation. Extensive experiments across seven datasets and three model architectures demonstrate our method's versatility and effectiveness. Its plug-and-play characteristic makes it compatible with a wide range of distillation frameworks.
****

MAGiC-SLAM: Multi-Agent Gaussian Globally Consistent SLAM
Vladimir Yugay, Theo Gevers, Martin R. Oswald; Proceedings of the Computer Visio n and Pattern Recognition Conference (CVPR), 2025, pp. 6741-6750
Simultaneous localization and mapping (SLAM) systems with novel view synthesis c apabilities are widely used in computer vision, with applications in augmented r eality, robotics, and autonomous driving. However, existing approaches are limit ed to single-agent operation. Recent work has addressed this problem using a dis tributed neural scene representation. Unfortunately, existing methods are slow, cannot accurately render real-world data, are restricted to two agents, and have limited tracking accuracy. In contrast, we propose a rigidly deformable 3D Gaus sian-based scene representation that dramatically speeds up the system. However, improving tracking accuracy and reconstructing a globally consistent map from m ultiple agents remains challenging due to trajectory drift and discrepancies acr oss agents' observations. Therefore, we propose new tracking and map-merging mec hanisms and integrate loop closure in the Gaussian-based SLAM pipeline. We evalu ate \ours on synthetic and real-world datasets and find it more accurate and fas ter than the state of the art.
****

Dispider: Enabling Video LLMs with Active Real-Time Interaction via Disentangled Perception, Decision, and Reaction
Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, Jiaqi Wang; Proceedings of the Computer Vision and Pattern Recognition Con ference (CVPR), 2025, pp. 24045-24055
Active Real-time interaction with video LLMs introduces a new paradigm for human -computer interaction, where the model not only understands user intent but also responds while continuously processing streaming video on the fly. Unlike offli ne video LLMs, which analyze the entire video before answering questions, active real-time interaction requires three capabilities: 1) Perception: real-time vid eo monitoring and interaction capturing. 2) Decision: raising proactive interact ion in proper situations, 3) Reaction: continuous interaction with users. Howeve r, inherent conflicts exist among the desired capabilities. The Decision and Rea ction require a contrary Perception scale and grain, and the autoregressive deco ding blocks the real-time Perception and Decision during the Reaction. To unify the conflicted capabilities within a harmonious system, we present Dispider, a s olution built on a Disentangled Perception, Decision, and Reaction framework. Di spider features a lightweight Proactive Streaming Video Processing module that t racks the video stream and identifies optimal moments for interaction. Once the interaction is triggered, an asynchronous Precise Interaction module provides de tailed responses, while the processing module continues to monitor the video in the meantime. Our disentangled and asynchronous design ensures timely, contextua lly accurate, and computationally efficient responses, making Dispider ideal for active real-time interaction for long-duration video streams. Experiments prove that Dispider outperforms existing methods not only in its superior understandi ng of video content in conventional video QA settings, but also in proactive res ponse capability and temporal awareness under the streaming setting.
****

NTClick: Achieving Precise Interactive Segmentation With Noise-tolerant Clicks
Chenyi Zhang, Ting Liu, Xiaochao Qu, Luoqi Liu, Yao Zhao, Yunchao Wei; Proceedin gs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8 921-8930
Interactive segmentation is a pivotal task in computer vision, focused on predic ting precise masks with minimal user input. Although the click has recently beco me the most prevalent form of interaction due to its flexibility and efficiency, its advantages diminish as the complexity and details of target objects increas e because it's time-consuming and user-unfriendly to precisely locate and click on narrow, fine regions. To tackle this problem, we propose NTClick, a powerful click-based interactive segmentation method capable of predicting accurate masks even with imprecise user clicks when dealing with intricate targets. We first i ntroduce a novel interaction form called Noist-tolerant Click, a type of click t hat does not require user's precise localization when selecting fine regions. Th

en, we design a two-stage workflow, consisting of an Explicit Coarse Perception network for initial estimation and a High Resolution Refinement network for final classification. Quantitative results across extensive datasets demonstrate that NTClick not only maintains an efficient and flexible interaction mode but also significantly outperforms existing methods in segmentation accuracy.
****

Show and Segment: Universal Medical Image Segmentation via In-Context Learning
Yunhe Gao, Di Liu, Zhuowei Li, Yunsheng Li, Dongdong Chen, Mu Zhou, Dimitris N. Metaxas; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20830-20840
Medical image segmentation remains challenging due to the vast diversity of anatomical structures, imaging modalities, and segmentation tasks. While deep learning has made significant advances, current approaches struggle to generalize as they require task-specific training or fine-tuning on unseen classes. We present Iris, a novel In-context Reference Image guided Segmentation framework that enables flexible adaptation to novel tasks through the use of reference examples without fine-tuning. At its core, Iris features a lightweight context task encoding module that distills task-specific information from reference context image-label pairs. This rich context embedding information is used to guide the segmentation of target objects. Given a decoupled architecture on 3D data processing, Iris supports diverse inference strategies including one-shot inference, context example ensemble, object-level context example retrieval, and in-context tuning. Through comprehensive evaluation across twelve datasets, we demonstrate that Iris performs strongly compared to specialized supervised models on in-distribution tasks. On seven held-out dataset, Iris shows superior generalization to out-of-distribution data and unseen classes. Further, Iris's task encoding module can automatically discover anatomical relationships across datasets and modalities, offering insights into cross-modality medical objects without explicit anatomical supervision.
****

MVGenMaster: Scaling Multi-View Generation from Any Image via 3D Priors Enhanced Diffusion Model
Chenjie Cao, Chaohui Yu, Shang Liu, Fan Wang, Xiangyang Xue, Yanwei Fu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6045-6056
We introduce MVGenMaster, a multi-view diffusion model enhanced with 3D priors to address versatile Novel View Synthesis (NVS) tasks. MVGenMaster leverages 3D priors that are warped using metric depth and camera poses, significantly enhancing both generalization and 3D consistency in NVS. Our model features a simple yet effective pipeline that can generate up to 100 novel views conditioned on variable reference views and camera poses with a single forward process. Additionally, we have developed a comprehensive large-scale multi-view image dataset called MvD-1M, comprising up to 1.6 million scenes, equipped with well-aligned metric depth to train MVGenMaster. Moreover, we present several training and model modifications to strengthen the model with scaled-up datasets. Extensive evaluations across in- and out-of-domain benchmarks demonstrate the effectiveness of our proposed method and data formulation.
****

CADCrafter: Generating Computer-Aided Design Models from Unconstrained Images
Cheng Chen, Jiacheng Wei, Tianrun Chen, Chi Zhang, Xiaofeng Yang, Shangzhan Zhang, Bingchen Yang, Chuan-Sheng Foo, Guosheng Lin, Qixing Huang, Fayao Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11073-11082
Creating CAD digital twins from the physical world is crucial for manufacturing, design, and simulation. However, current methods typically rely on costly 3D scanning with labor-intensive post-processing. To provide a user-friendly design process, we explore the problem of reverse engineering from unconstrained real-world CAD images that can be easily captured by users of all experiences. However, the scarcity of real-world CAD data poses challenges in directly training such models. To tackle these challenges, we propose CADCrafter, an image-to-parametri

c CAD model generation framework that trains solely on synthetic textureless CAD data while testing on real-world images. To bridge the significant representation disparity between images and parametric CAD models, we introduce a geometry encoder to accurately capture diverse geometric features. Moreover, the texture-invariant properties of the geometric features can also facilitate the generalization to real-world scenarios. Since compiling CAD parameter sequences into explicit CAD models is a non-differentiable process, the network training inherently lacks explicit geometric supervision. To impose geometric validity constraints, we employ direct preference optimization (DPO) to fine-tune our model with the automatic code checker feedback on CAD sequence quality. Furthermore, we collected a real-world dataset, comprised of multi-view images and corresponding CAD command sequence pairs, to evaluate our method. Experimental results demonstrate that our approach can robustly handle real unconstrained CAD images, and even generalize to unseen general objects.

****

Bayesian Test-Time Adaptation for Vision-Language Models

Lihua Zhou, Mao Ye, Shuaifeng Li, Nianxin Li, Xiatian Zhu, Lei Deng, Hongbin Liu, Zhen Lei; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29999-30009

Test-time adaptation with pre-trained vision-language models, such as CLIP, aims to adapt the model to new, potentially out-of-distribution test data. Existing methods calculate the similarity between visual embedding and learnable class embeddings, which are initialized by text embeddings, for zero-shot image classification. In this work, we first analyze this process based on Bayes theorem, and observe that the core factors influencing the final prediction are the likelihood and the prior. However, existing methods essentially focus on adapting class embeddings to adapt likelihood, but they often ignore the importance of prior. To address this gap, we propose a novel approach, Bayesian Class Adaptation (BCA), which in addition to continuously updating class embeddings to adapt likelihood, also uses the posterior of incoming samples to continuously update the prior for each class embedding. This dual updating mechanism allows the model to better adapt to distribution shifts and achieve higher prediction accuracy. Our method not only surpasses existing approaches in terms of performance metrics but also maintains superior inference rates and memory usage, making it highly efficient and practical for real-world applications.

****

Generative Multiview Relighting for 3D Reconstruction under Extreme Illumination Variation

Hadi Alzayer, Philipp Henzler, Jonathan T. Barron, Jia-Bin Huang, Pratul P. Srinivasan, Dor Verbin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10933-10942

Reconstructing the geometry and appearance of objects from photographs taken in different environments is difficult as the illumination and therefore the object appearance vary across captured images. This is particularly challenging for more specular objects whose appearance strongly depends on the viewing direction. Some prior approaches model appearance variation across images using a per-image embedding vector, while others use physically-based rendering to recover the materials and per-image illumination. Such approaches fail at faithfully recovering view-dependent appearance given significant variation in input illumination and tend to produce mostly diffuse results. We present an approach that reconstructs objects from images taken under different illuminations by first relighting the images under a single reference illumination with a multiview relighting diffusion model and then reconstructing the object's geometry and appearance with a radiance field architecture that is robust to the small remaining inconsistencies among the relit images. We validate our proposed approach on both simulated and real datasets and demonstrate that it greatly outperforms existing techniques at reconstructing high-fidelity appearance from images taken under extreme illumination variation. Moreover, our approach is particularly effective at recovering view-dependent "shiny" appearance which cannot be reconstructed by prior methods.

****
Causal Composition Diffusion Model for Closed-loop Traffic Generation
Haohong Lin, Xin Huang, Tung Phan, David Hayden, Huan Zhang, Ding Zhao, Siddhartha Srinivasa, Eric Wolff, Hongge Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27542-27552
Simulation is critical for safety evaluation in autonomous driving, particularly in capturing complex interactive behaviors. However, generating **realistic** and **controllable** traffic scenarios in long-tail situations remains a significant challenge. Existing generative models suffer from the conflicting objective between user-defined controllability and realism constraints, which is amplified in safety-critical contexts. In this work, we introduce the **C**ausal **C**ompositional **Diff**usion Model (***CCDiff***), a structure-guided diffusion framework to address these challenges. We first formulate the learning of controllable and realistic closed-loop simulation as a constrained optimization problem. Then, CCDiff maximizes controllability while adhering to realism by automatically identifying and injecting causal structures directly into the diffusion process, providing structured guidance to enhance both realism and controllability. Through rigorous evaluations on benchmark datasets and in a closed-loop simulator, CCDiff demonstrates substantial gains over state-of-the-art approaches in generating realistic and user-preferred trajectories. Our results show CCDiff's effectiveness in extracting and leveraging causal structures, showing improved closed-loop performance based on key metrics such as collision rate, off-road rate, FDE, and comfort. For more details, welcome to check our project website: https://sites.google.com/view/ccdiff/.
****
Change3D: Revisiting Change Detection and Captioning from A Video Modeling Perspective
Duowang Zhu, Xiaohu Huang, Haiyan Huang, Hao Zhou, Zhenfeng Shao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24011-24022
In this paper, we present Change3D, a framework that reconceptualizes the change detection and captioning tasks through video modeling. Recent methods have achieved remarkable success by regarding each pair of bi-temporal images as separate frames. They employ a shared-weight image encoder to extract spatial features and then use a change extractor to capture differences between the two images. However, image feature encoding, being a task-agnostic process, cannot attend to changed regions effectively. Furthermore, different change extractors designed for various change detection and captioning tasks make it difficult to have a unified framework. To tackle these challenges, Change3D regards the bi-temporal images as comprising two frames akin to a tiny video. By integrating learnable perception frames between the bi-temporal images, a video encoder enables the perception frames to interact with the images directly and perceive their differences. Therefore, we can get rid of the intricate change extractors, providing a unified framework for different change detection and captioning tasks. We verify Change3D on multiple tasks, encompassing change detection (including binary change detection, semantic change detection, and building damage assessment) and change captioning, across eight standard benchmarks. Without bells and whistles, this simple yet effective framework can achieve superior performance with an ultra-light video model comprising only ~6%-13% of the parameters and ~8%-34% of the FLOPs compared to state-of-the-art methods. We hope that Change3D could be an alternative to 2D-based models and facilitate future research.
****
DyMO: Training-Free Diffusion Model Alignment with Dynamic Multi-Objective Scheduling
Xin Xie, Dong Gong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13220-13230
Text-to-image diffusion model alignment is critical for improving the alignment between the generated images and human preferences. While training-based methods are constrained by high computational costs and dataset requirements, training-free alignment methods remain underexplored and are often limited by inaccurate

guidance. We propose a plug-and-play training-free alignment method, DyMO, for a ligning the generated images and human preferences during inference. Apart from text-aware human preference scores, we introduce a semantic alignment objective for enhancing the semantic alignment in the early stages of diffusion, relying o n the fact that the attention maps are effective reflections of the semantics in noisy images. We propose dynamic scheduling of multiple objectives and intermed iate recurrent steps to reflect the requirements at different steps. Experiments with diverse pre-trained diffusion models and metrics demonstrate the effective ness and robustness of the proposed method.
****

HiMoR: Monocular Deformable Gaussian Reconstruction with Hierarchical Motion Rep resentation
Yiming Liang, Tianhan Xu, Yuta Kikuchi; Proceedings of the Computer Vision and P attern Recognition Conference (CVPR), 2025, pp. 886-895
We present Hierarchical Motion Representation (HiMoR), a novel deformation repre sentation for 3D Gaussian primitives capable of achieving high-quality monocular dynamic 3D reconstruction. The insight behind HiMoR is that motions in everyday scenes can be decomposed into coarser motions that serve as the foundation for finer details. Using a tree structure, HiMoR's nodes represent different levels of motion detail, with shallower nodes modeling coarse motion for temporal smoot hness and deeper nodes capturing finer motion. Additionally, our model uses a fe w shared motion bases to represent motions of different sets of nodes, aligning with the assumption that motion tends to be smooth and simple. This motion repre sentation design provides Gaussians with a more structured deformation, maximizi ng the use of temporal relationships to tackle the challenging task of monocular dynamic 3D reconstruction. We also propose using a more reliable perceptual met ric as an alternative, given that pixel-level metrics for evaluating monocular d ynamic 3D reconstruction can sometimes fail to accurately reflect the true quali ty of reconstruction. Extensive experiments demonstrate our method's efficacy in achieving superior novel view synthesis from challenging monocular videos with complex motions.
****

GENIUS: A Generative Framework for Universal Multimodal Search
Sungyeon Kim, Xinliang Zhu, Xiaofan Lin, Muhammet Bastan, Douglas Gray, Suha Kwa k; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19659-19669
Generative retrieval is an emerging approach in information retrieval that gener ates identifiers (IDs) of target data based on a query, providing an efficient a lternative to traditional embedding-based retrieval methods. However, existing m odels are task-specific and fall short of embedding-based retrieval in performan ce. This paper proposes GENIUS, a universal generative retrieval framework suppo rting diverse tasks across multiple modalities and domains. At its core, GENIUS introduces modality-decoupled semantic quantization, transforming multimodal dat a into discrete IDs encoding both modality and semantics. Moreover, to enhance g eneralization, we propose a query augmentation that interpolates between a query and its target, allowing GENIUS to adapt to varied query forms. Evaluated on th e M-BEIR benchmark, it surpasses prior generative methods by a clear margin. Unl ike embedding-based retrieval, GENIUS consistently maintains high retrieval spee d across database size, with competitive performance across multiple benchmarks. With additional re-ranking, GENIUS often achieves results close to those of emb edding-based methods while preserving efficiency.
****

Enhanced Visual-Semantic Interaction with Tailored Prompts for Pedestrian Attrib ute Recognition
Junyi Wu, Yan Huang, Min Gao, Yuzhen Niu, Yuzhong Chen, Qiang Wu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9570-9 579
Pedestrian attribute recognition (PAR) seeks to predict multiple semantic attrib utes associated with a specific pedestrian. There are two types of approaches fo r PAR: unimodal framework and bimodal framework. The former one is to seek a rob

ust visual feature. However, the lack of exploiting semantic feature of linguist ic modality is the main concern. The latter one utilizes prompt learning techniq ues to integrate linguistic data. However, static prompt templates and simple bi modal concatenation cannot to capture the extensive intra-class attribute variab ility and support active modalities collaboration. In this paper, we propose an Enhanced Visual-Semantic Interaction with Tailored Prompts (EVSITP) framework fo r PAR. We present an Image-Conditional Dual-Prompt Initialization Module (IDIM) to adaptively generate context-sensitive prompts from visual inputs. Subsequentl y, a Prompt Enhanced and Regularization Module (PERM) is proposed to strengthen linguistic information from IDIM. We further design a Bimodal Mutual Interaction Module (BMIM) to ensure bidirectional modalities communication. In addition, e xisting PAR datasets are collected over a short period in limited scenarios, whi ch do not align with real-world scenarios. Therefore, we annotate a long-term pe rson re-identification dataset to create a new PAR dataset, Celeb-PAR. Experimen ts on several challenging PAR datasets show that our method outperforms state-of -the-art approaches.

****

SF3D: Stable Fast 3D Mesh Reconstruction with UV-unwrapping and Illumination Dis entanglement

Mark Boss, Zixuan Huang, Aaryaman Vasishta, Varun Jampani; Proceedings of the Co mputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16240-16250

We present SF3D, a novel method for rapid and high-quality textured object mesh reconstruction from a single image in just 0.5 seconds. Unlike most existing app roaches, SF3D is explicitly trained for mesh generation, incorporating a fast UV unwrapping technique that enables swift texture generation rather than relying on vertex colors. The method also learns to predict material parameters and norm al maps to enhance the visual quality of the reconstructed 3D meshes. Furthermor e, SF3D integrates a delighting step to effectively remove low-frequency illumin ation effects, ensuring that the reconstructed meshes can be easily used in nove l illumination conditions. Experiments demonstrate the superior performance of S F3D over the existing techniques

****

HSI-GPT: A General-Purpose Large Scene-Motion-Language Model for Human Scene Int eraction

Yuan Wang, Yali Li, Xiang Li, Shengjin Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7147-7157

While flourishing developments have been witnessed in text-to-motion generation, synthesizing physically realistic, controllable, language-conditioned Human Sce ne Interactions (HSI) remains a relatively underexplored landscape. Current HSI methods naively rely on conditional Variational AutoEncoder (cVAE) and diffusion models. They are typically associated with limited modalities of control signal s and task-specific frameworks design, leading to inflexible adaptation across v arious interaction scenarios and descriptive-unfaithful motions in diverse 3D ph ysical environments. In this paper, we propose HSI-GPT, a General-Purpose Large Scene-Motion-Language Model that applies "next-token prediction" paradigm of Lar ge Language Models to the HSI domain. HSI-GPT not only exhibits remarkable flexi bility to accommodate diverse control signals (3D scenes, textual commands, key- frame poses, as well as scene affordances), but it seamlessly supports various H SI-related tasks (e.g., multi-modal controlled HSI generation, HSI understanding , and general motion completion in 3D scenes). First, HSI-GPT quantizes textual descriptions and human motions into discrete, LLM-interpretable tokens with mult i-modal tokenizers. Inspired by multi-modal learning, we develop a recipe for al igning mixed-modality tokens into the shared embedding space of LLMs. These inte raction tokens are then organized into unified instruction following prompts, al lowing HSI-GPT to fine-tune on question-and-answer tasks. Extensive experiments and visualizations validate that our general-purpose HSI-GPT model delivers exce ptional performance across multiple HSI-related tasks.

****

Towards Precise Embodied Dialogue Localization via Causality Guided Diffusion

Haoyu Wang, Le Wang, Sanping Zhou, Jingyi Tian, Zheng Qin, Yabing Wang, Gang Hua

, Wei Tang; Proceedings of the Computer Vision and Pattern Recognition Conferenc e (CVPR), 2025, pp. 13350-13360
Embodied localization based on vision and natural language dialogues presents a persistent challenge in embodied intelligence. Existing methods often approach t his task as an image translation problem, leveraging encoder-decoder architectur es to predict heatmaps. However, these methods frequently experience a deficienc y in accuracy, largely due to their heavy reliance on resolution. To address thi s issue, we introduce CGD, a novel framework that utilizes causality guided diff usion model to directly model coordinate distributions. Specifically, CGD employ s a denoising network to regress coordinates, while integrating causal learning modules, namely back-door adjustment (BDA) and front-door adjustment (FDA) to mi tigate confounders during the diffusion process. This approach reduces the depen dency on high resolution for improving accuracy, while effectively minimizing sp urious correlations, thereby promoting unbiased learning. By guiding the denoisi ng process with causal adjustments, CGD offers flexible control over intensity, ensuring seamless integration with diffusion models. Experimental results demons trate that CGD outperforms state-of-the-art methods across all metrics. Addition ally, we also evaluate CGD in a multi-shot setting, achieving consistently high accuracy.
****

Vid2Avatar-Pro: Authentic Avatar from Videos in the Wild via Universal Prior
Chen Guo, Junxuan Li, Yash Kant, Yaser Sheikh, Shunsuke Saito, Chen Cao; Proceed ings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5559-5570
We present Vid2Avatar-Pro, a method to create photorealistic and animatable 3D h uman avatars from monocular in-the-wild videos. Building a high-quality avatar t hat supports animation with diverse poses from a monocular video is challenging because the observation of pose diversity and view points is inherently limited. The lack of pose variations typically leads to poor generalization to novel pos es, and avatars can easily overfit to limited input view points, producing artif acts and distortions from other views. In this work, we address these limitation s by leveraging a universal prior model (UPM) learned from a large corpus of mul ti-view clothed human performance capture data. We build our representation on t op of expressive 3D Gaussians with canonical front and back maps shared across i dentities. Once the UPM is learned to accurately reproduce the large-scale multi -view human images, we fine-tune the model with an in-the-wild video via inverse rendering to obtain a personalized photorealistic human avatar that can be fait hfully animated to novel human motions and rendered from novel views. The experi ments show that our approach based on the learned universal prior sets a new sta te-of-the-art in monocular avatar reconstruction by substantially outperforming existing approaches relying only on heuristic regularization or a shape prior of minimally clothed bodies (e.g., SMPL) on publicly available datasets.
****

RoomPainter: View-Integrated Diffusion for Consistent Indoor Scene Texturing
Zhipeng Huang, Wangbo Yu, Xinhua Cheng, Chengshu Zhao, Yunyang Ge, Mingyi Guo, L i Yuan, Yonghong Tian; Proceedings of the Computer Vision and Pattern Recognitio n Conference (CVPR), 2025, pp. 574-584
Indoor scene texture synthesis has garnered significant interest due to its impo rtant potential applications in virtual reality, digital media and creative arts . Existing diffusion-model-based researches either rely on per-view inpainting t echniques, which are plagued by severe cross-view inconsistencies and conspicuou s seams, or adopt optimization-based approaches that involve substantial computa tional overhead. In this work, we present **RoomPainter**, a framework that seam lessly integrates efficiency and consistency to achieve high-fidelity texturing of indoor scenes. The core of RoomPainter features a zero-shot technique that ef fectively adapts a 2D diffusion model for 3D-consistent texture synthesis, along with a two-stage generation strategy that ensures both global and local consist ency. Specifically, we introduce Attention-Guided Multi-View Integrated Sampling (**MVIS**) combined with a neighbor-integrated attention mechanism for zero-sho t texture map generation. Using the **MVIS**, we firstly generate texture map fo

r the entire room to ensure global consistency, then adopt its variant, namely A ttention-Guided Multi-View Integrated Repaint Sampling (**MVRS**) to repaint ind ividual instances within the room, thereby further enhancing local consistency a nd addressing the occlusion problem. Experiments demonstrate that RoomPainter ac hieves superior performance for indoor scene texture synthesis in visual quality , global consistency and generation efficiency.
****

Attribute-formed Class-specific Concept Space: Endowing Language Bottleneck Mode l with Better Interpretability and Scalability
Jianyang Zhang, Qianli Luo, Guowu Yang, Wenjing Yang, Weide Liu, Guosheng Lin, F engmao Lv; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30291-30300
Language Bottleneck Models (LBMs) are proposed to achieve interpretable image re cognition by classifying images based on textual concept bottlenecks. However, c urrent LBMs simply list all concepts together as the bottleneck layer, leading t o the spurious cue inference problem and cannot generalized to unseen classes. T o address these limitations, we propose the Attribute-formed Language Bottleneck Model (ALBM). ALBM organizes concepts in the attribute-formed class-specific sp ace, where concepts are descriptions of specific attributes for specific classes . In this way, ALBM can avoid the spurious cue inference problem by classifying solely based on the essential concepts of each class. In addition, the cross-cla ss unified attribute set also ensures that the concept spaces of different class es have strong correlations, as a result, the learned concept classifier can be easily generalized to unseen classes. Moreover, to further improve interpretabil ity, we propose Visual Attribute Prompt Learning (VAPL) to extract visual featur es on fine-grained attributes. Furthermore, to avoid labor-intensive concept ann otation, we propose the Description, Summary, and Supplement (DSS) strategy to a utomatically generate high-quality concept sets with a complete and precise attr ibute. Extensive experiments on 9 widely used few-shot benchmarks demonstrate th e interpretability, transferability, and performance of our approach. The code a nd collected concept sets are available at https://github.com/tiggers23/ALBM.
****

Customized Condition Controllable Generation for Video Soundtrack
Fan Qi, Kunsheng Ma, Changsheng Xu; Proceedings of the Computer Vision and Patte rn Recognition Conference (CVPR), 2025, pp. 23914-23924
Recent advances in latent diffusion models (LDMs) have enabled data-driven parad igms for video soundtrack generation, improving multimodal alignment capabilitie s. However, current two-stage frameworks--which separately optimize audio-visual correspondence and conditional audio synthesis--fundamentally limit joint model ing of dynamic acoustic properties. In this paper, we propose a novel framework for generating video soundtracks that simultaneously produces music and sound ef fect tailored to the video content. Our method incorporates a Contrastive Visual -Sound-Music pretraining process that maps these modalities into a unified featu re space, enhancing the model's ability to capture intricate audio dynamics. We design Spectrum Divergence Masked Attention for Unet to differentiate between th e unique characteristics of sound effect and music. We utilize Score-guided Nois e Iterative Optimization to provide musicians with customizable control during t he generation process. Extensive evaluations on the FilmScoreDB and SymMV&HIMV d atasets demonstrate that our approach significantly outperforms state-of-the-art baselines in both subjective and objective assessments, highlighting its potent ial as a robust tool for video soundtrack generation.
****

ProjAttacker: A Configurable Physical Adversarial Attack for Face Recognition vi a Projector
Yuanwei Liu, Hui Wei, Chengyu Jia, Ruqi Xiao, Weijian Ruan, Xingxing Wei, Joey T ianyi Zhou, Zheng Wang; Proceedings of the Computer Vision and Pattern Recogniti on Conference (CVPR), 2025, pp. 21248-21257
Previous physical adversarial attacks have shown that carefully crafted perturba tions can deceive face recognition systems, revealing critical security vulnerab ilities. However, these attacks often struggle to impersonate multiple targets a

nd frequently fail to bypass liveness detection. For example, attacks using huma n-skin masks are challenging to fabricate, inconvenient to swap between users, a nd often fail liveness detection due to facial occlusions. A projector, however, can generate content-rich light without obstructing the face, making it ideal f or non-intrusive attacks. Thus, we propose a novel physical adversarial attack u sing a projector and explore the superposition of projected and natural light to create adversarial facial images. This approach eliminates the need for physica l artifacts on the face, effectively overcoming these limitations. Specifically, our proposed ProjAttacker generates adversarial 3D textures that are projected onto human faces. To ensure physical realizability, we introduce a light reflect ion function that models complex optical interactions between projected light an d human skin, accounting for reflection and diffraction effects. Furthermore, we incorporate camera Image Signal Processing (ISP) simulation to maintain the rob ustness of adversarial perturbations across real-world diverse imaging condition s. Comprehensive evaluations conducted in both digital and physical scenarios va lidate the effectiveness of our method.

****

EfficientViM: Efficient Vision Mamba with Hidden State Mixer based State Space D uality

Sanghyeok Lee, Joonmyung Choi, Hyunwoo J. Kim; Proceedings of the Computer Visio n and Pattern Recognition Conference (CVPR), 2025, pp. 14923-14933

For the deployment of neural networks in resource-constrained environments, prio r works have built lightweight architectures with convolution and attention for capturing local and global dependencies, respectively. Recently, the state space model (SSM) has emerged as an effective operation for global interaction with i ts favorable linear computational cost in the number of tokens. To harness the e fficacy of SSM, we introduce Efficient Vision Mamba (EfficientViM), a novel arch itecture built on hidden state mixer-based state space duality (HSM-SSD) that ef ficiently captures global dependencies with further reduced computational cost. With the observation that the runtime of the SSD layer is driven by the linear p rojections on the input sequences, we redesign the original SSD layer to perform the channel mixing operation within compressed hidden states in the HSM-SSD lay er. Additionally, we propose multi-stage hidden state fusion to reinforce the re presentation power of hidden states and provide the design to alleviate the bott leneck caused by the memory-bound operations. As a result, the EfficientViM fami ly achieves a new state-of-the-art speed-accuracy trade-off on ImageNet-1k, offe ring up to a 0.7% performance improvement over the second-best model SHViT with faster speed. Further, we observe significant improvements in throughput and acc uracy compared to prior works, when scaling images or employing distillation tra ining. Code is available at https://github.com/mlvlab/EfficientViM.

****

A4A: Adapter for Adapter Transfer via All-for-All Mapping for Cross-Architecture Models

Keyu Tu, Mengqi Huang, Zhuowei Chen, Zhendong Mao; Proceedings of the Computer V ision and Pattern Recognition Conference (CVPR), 2025, pp. 18476-18485

Large-scale text-to-image models evolve rapidly in size and architecture. The ex isting adapters struggle to keep pace with these models, requiring extensive ret raining. This paper proposes a novel adapter transfer framework, A4A (Adapter fo r Adapter), which uses an all-for-all mapping approach to seamlessly transfer at tention-based adapters across different model architectures (e.g., U-Net to tran sformer). The framework consists of Coupling Space Projection and Upgraded Space Mapping. During Coupling Space Projection, all attention features of the pretra ined adapter are aggregated to fully capture the coupling relationship before be ing projected into a unified space. The unified space maintains coupling feature s in a consistent dimension, effectively and efficiently addressing feature scal e discrepancies arising from the base model's architecture. In the Upgraded Spac e Mapping Module, randomly initialized learnable features are introduced to conn ect the unified and upgraded spaces by integrating reference features via the at tention mechanism. The learned features are adaptively injected into the upgrade model through the Alignment module, which bridges the discrepancies between the

models using the all-for-all mapping. Experimental results on personalized image generation tasks demonstrate that A4A outperforms previous methods in transferring adapters while being the first to achieve adapter transfer across model architectures.

****

ViCaS: A Dataset for Combining Holistic and Pixel-level Video Understanding using Captions with Grounded Segmentation

Ali Athar, Xueqing Deng, Liang-Chieh Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19023-19035

Recent advances in multimodal large language models (MLLMs) have expanded research in video understanding, primarily focusing on high-level tasks such as video captioning and question-answering. Meanwhile, a smaller body of work addresses dense, pixel-precise segmentation tasks, which typically involve category-guided or referral-based object segmentation. Although both research directions are essential for developing models with human-level video comprehension, they have largely evolved separately, with distinct benchmarks and architectures. This paper aims to unify these efforts by introducing ViCaS, a new dataset containing thousands of challenging videos, each annotated with detailed, human-written captions and temporally consistent, pixel-accurate masks for multiple objects with phrase grounding. Our benchmark evaluates models on both holistic/high-level understanding and language-guided, pixel-precise segmentation. We also present carefully validated evaluation measures and propose an effective model architecture that can tackle our benchmark. Project page: https://ali2500.github.io/vicas-project/

****

A Universal Scale-Adaptive Deformable Transformer for Image Restoration across Diverse Artifacts

Xuyi He, Yuhui Quan, Ruotao Xu, Hui Ji; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12731-12741

Structured artifacts are semi-regular, repetitive patterns that closely intertwine with genuine image content, making their removal highly challenging. In this paper, we introduce the Scale-Adaptive Deformable Transformer, an network architecture specifically designed to eliminate such artifacts from images. The proposed network features two key components: a scale-enhanced deformable convolution module for modeling scale-varying patterns with abundant orientations and potential distortions, and a scale-adaptive deformable attention mechanism for capturing long-range relationships among repetitive patterns with different sizes and non-uniform spatial distributions. Extensive experiments show that our network consistently outperforms state-of-the-art methods in diverse artifact removal tasks, including image deraining, image demoireing, and image debanding.

****

WISE: A Framework for Gigapixel Whole-Slide-Image Lossless Compression

Yu Mao, Jun Wang, Nan Guan, Chun Jason Xue; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29342-29351

Whole-Slide Images (WSIs) have revolutionized medical analysis by presenting high-resolution images of the whole tissue slide. Despite avoiding the physical storage of the slides, WSIs require considerable data volume, which makes the storage and maintenance of WSI records costly and unsustainable. To this end, this work presents the first investigation of lossless compression of WSI images. Interestingly, we find that most existing compression methods fail to compress the WSI images effectively. Furthermore, our analysis reveals that the failure of existing compressors is mainly due to information irregularity in WSI images. To resolve this issue, we develop a simple yet effective lossless compressor called WISE, specifically designed for WSI images. WISE employs a hierarchical encoding strategy to extract effective bits, reducing the entropy of the image and then adopting a dictionary-based method to handle the irregular frequency patterns. Through extensive experiments, we show that WISE can effectively compress the gigapixel WSI images to 36 times on average and up to 136 times.

****

Gromov-Wasserstein Problem with Cyclic Symmetry

Shoichiro Takeda, Yasunori Akagi; Proceedings of the Computer Vision and Pattern

Recognition Conference (CVPR), 2025, pp. 21011-21020
We propose novel fast algorithms for the Gromov--Wasserstein problem (GW) with cyclic symmetry of input data. This problem naturally appears as an object-matching task, which underlies various real-world computer vision applications, e.g., image registration, point cloud registration, stereo matching, and 3D reconstruction. Gradient-based algorithms have been widely used to solve GW, and our main idea is to utilize the following remarkable property that emerges in GW with cyclic symmetry: By setting the initial solution to have cyclic symmetry, all intermediate solutions and matrices that appear in the gradient-based algorithms have the same cyclic symmetry until convergence. Based on this property, our gradient-based algorithms restrict the solution space to have cyclic symmetry and update only one symmetric part of solutions and matrices at each iteration, resulting in faster computation. Moreover, our algorithms solve the optimal transport problem at each iteration, which also exhibits cyclic symmetry. This problem can be solved efficiently, and as a result, our algorithms perform significantly faster. Experiments showed the effectiveness of our algorithms in synthetic and real-world data with strict and approximate cyclic symmetry.

****

IRIS: Inverse Rendering of Indoor Scenes from Low Dynamic Range Images
Chih-Hao Lin, Jia-Bin Huang, Zhengqin Li, Zhao Dong, Christian Richardt, Tuotuo Li, Michael Zollhöfer, Johannes Kopf, Shenlong Wang, Changil Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 465-474
Inverse rendering seeks to recover 3D geometry, surface material, and lighting from captured images, enabling advanced applications such as novel-view synthesis, relighting, and virtual object insertion. However, most existing techniques rely on high dynamic range (HDR) images as input, limiting accessibility for general users. In response, we introduce IRIS, an inverse rendering framework that recovers the physically based material, spatially-varying HDR lighting, and camera response functions from multi-view, low-dynamic-range (LDR) images. By eliminating the dependence on HDR input, we make inverse rendering technology more accessible. We evaluate our approach on real-world and synthetic scenes and compare it with state-of-the-art methods. Our results show that IRIS effectively recovers HDR lighting, accurate material, and plausible camera response functions, supporting photorealistic relighting and object insertion.

****

SimAvatar: Simulation-Ready Avatars with Layered Hair and Clothing
Xueting Li, Ye Yuan, Shalini De Mello, Gilles Daviet, Jonathan Leaf, Miles Macklin, Jan Kautz, Umar Iqbal; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26320-26330
We introduce SimAvatar, a framework designed to generate simulation-ready clothed 3D human avatars from a text prompt. Current text-driven human avatar generation methods either model hair, clothing and human body using a unified geometry or produce hair and garments that are not easily adaptable for simulation within existing graphics pipelines. The primary challenge lies in representing the hair and garment geometry in a way that allows leveraging established prior knowledge from foundational image diffusion models (e.g., Stable Diffusion) while being simulation-ready using either physics or neural simulators. To address this task, we propose a two-stage framework that combines the flexibility of 3D Gaussians with simulation-ready hair strands and garment meshes. Specifically, we first leverage two text-conditioned diffusion models to generate garment mesh and hair strands from the given text prompt. To leverage prior knowledge from foundational diffusion models, we attach 3D Gaussians to the body mesh, garment mesh, as well as hair strands and learn the avatar appearance through optimization. To drive the avatar given a pose sequence, we first apply physics simulators onto the garment meshes and hair strands. We then transfer the motion onto 3D Gaussians through carefully designed mechanism for different body parts. As a result, our synthesized avatars have vivid texture and realistic dynamic motion. To the best of our knowledge, our method is the first to produce highly realistic, fully simulation-ready 3D avatars, surpassing the capabilities of current approaches.

****
Test-Time Backdoor Detection for Object Detection Models
Hangtao Zhang, Yichen Wang, Shihui Yan, Chenyu Zhu, Ziqi Zhou, Linshan Hou, Shengshan Hu, Minghui Li, Yanjun Zhang, Leo Yu Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24377-24386
Object detection models are vulnerable to backdoor attacks, where attackers poison a small subset of training samples by embedding a predefined trigger to manipulate prediction. Detecting poisoned samples (i.e., those containing triggers) at test time can prevent backdoor activation. However, unlike image classification tasks, the unique characteristics of object detection---particularly its output of numerous objects---pose fresh challenges for backdoor detection. The complex attack effects (e.g., "ghost" object emergence or "vanishing" object) further render current defenses fundamentally inadequate. To this end, we design TRAnsformation Consistency Evaluation (TRACE), a brand-new method for detecting poisoned samples at test time in object detection. Our journey begins with two intriguing observations: (1) poisoned samples exhibit significantly more consistent detection results than clean ones across varied backgrounds. (2) clean samples show higher detection consistency when introduced to different focal information. Based on these phenomena, TRACE applies foreground and background transformations to each test sample, then assesses transformation consistency by calculating the variance in objects confidences. TRACE achieves black-box, universal backdoor detection, with extensive experiments showing a 30% improvement in AUROC over state-of-the-art defenses and resistance to adaptive attacks.
****
Towards Precise Scaling Laws for Video Diffusion Transformers
Yuanyang Yin, Yaqi Zhao, Mingwu Zheng, Ke Lin, Jiarong Ou, Rui Chen, Victor Shea-Jay Huang, Jiahao Wang, Xin Tao, Pengfei Wan, Di Zhang, Baoqun Yin, Wentao Zhang, Kun Gai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18155-18165
Achieving optimal performance of video diffusion transformers within given data and compute budget is crucial due to their high training costs. This necessitates precisely determining the optimal model size and training hyperparameters before large-scale training. While scaling laws are employed in language models to predict performance, their existence and accurate derivation in visual generation models remain underexplored. In this paper, we systematically analyze scaling laws for video diffusion transformers and confirm their presence. Moreover, we discover that, unlike language models, video diffusion models are more sensitive to learning rate and batch size--two hyperparameters often not precisely modeled. To address this, we propose a new scaling law that predicts optimal hyperparameters for any model size and compute budget. Under these optimal settings, we achieve comparable performance and reduce inference costs by 40.1% compared to conventional scaling methods, within a compute budget of 1e10 TFlops. Furthermore, we establish a more generalized and precise relationship among validation loss, any model size, and compute budget. This enables performance prediction for non-optimal model sizes, which may also be appealed under practical inference cost constraints, achieving a better trade-off.
****
RoGSplat: Learning Robust Generalizable Human Gaussian Splatting from Sparse Multi-View Images
Junjin Xiao, Qing Zhang, Yonewei Nie, Lei Zhu, Wei-Shi Zheng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5980-5990
This paper presents RoGSplat, a novel approach for synthesizing high-fidelity novel views of unseen human from sparse multi-view images, while requiring no cumbersome per-subject optimization. Unlike previous methods that typically struggle with sparse views with few overlappings and are less effective in reconstructing complex human geometry, the proposed method enables robust reconstruction in such challenging conditions. Our key idea is to lift SMPL vertices to dense and reliable 3D prior points representing accurate human body geometry, and then regress human Gaussian parameters based on the points. To account for possible misalignment between SMPL model and images, we propose to predict image-aligned 3D pr

ior points by leveraging both pixel-level features and voxel-level features, fro m which we regress the coarse Gaussians. To enhance the ability to capture high- frequency details, we further render depth maps from the coarse 3D Gaussians to help regress fine-grained pixel-wise Gaussians. Experiments on several benchmark datasets demonstrate that our method outperforms state-of-the-art methods in no vel view synthesis and cross-dataset generalization. Our code is available at ht tps://github.com/iSEE-Laboratory/RoGSplat.
****

SDBF: Steep-Decision-Boundary Fingerprinting for Hard-Label Tampering Detection of DNN Models

Xiaofan Bai, Shixin Li, Xiaojing Ma, Bin Benjamin Zhu, Dongmei Zhang, Linchen Yu ; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29278-29287

Cloud-based AI systems offer significant benefits but also introduce vulnerabili ties, making deep neural network (DNN) models susceptible to malicious tampering . This tampering may involve harmful behavior injection or resource reduction, c ompromising model integrity and performance. To detect model tampering, hard-lab el fingerprinting techniques generate sensitive samples to probe and reveal tamp ering. Existing fingerprinting methods are mainly based on gradient-defined sens itivity decision boundary, with the latter showing a manifest superior detection performance. However, all existing fingerprinting methods either suffer from in sufficient sensitivity or incur high computational costs. In this paper, we theo retically analyze the black-box co-optimal tampering detection sensitivity of fi ngerprint samples in the context of decision boundary and gradient-defined sensi tivity. Based on this, we further propose Steep-Decision-Boundary Fingerprinting (SDBF), a novel lightweight approach for hard-label tampering detection that in herently and efficiently combines the strengths of existing fingerprinting techn iques. SDBF places fingerprint samples near the steep decision boundary, where t he outputs of samples are inherently highly sensitive to tampering. We also desi gn a Max Boundary Coverage Strategy (MBCS), which enhances samples' diversity o ver the decision boundary. Theoretical analysis and extensive experimental resul ts show that SDBF outperforms existing SOTA hard-label fingerprinting methods in both sensitivity and efficiency.
****

EnliveningGS: Active Locomotion of 3DGS

Siyuan Shen, Tianjia Shao, Kun Zhou, Chenfanfu Jiang, Yin Yang; Proceedings of t he Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 896-905

This paper presents a novel pipeline named EnliveningGS, which enables active lo comotion of 3D models represented with 3D Gaussian splatting (3DGS). We are insp ired by the fact that real-world lives pose their bodies in a natural and physic ally meaningful manner by compressing or elongating muscle fibers embedded in th e body. EnliveningGS aims to replicate the similar functionality of 3DGS models so that the object within a 3DGS scene acts like a living creature rather than a static shape --- they walk, jump, and twist in the scene under provided motion trajectories driven by muscle activations. While the concept is straightforward, many challenging technical difficulties need to be taken care of. Synthesizing realistic locomotion of a 3DGS model embodies an inverse physics problem of very high dimensions. The core challenge is how to efficiently and robustly model fr ictional contacts between an "enlivened model" and the environment, as it is the composition of contact/collision/friction forces triggered by muscle activation that generates the final movement of the object. We propose a hybrid numerical method mixing LCP and penalty method to tackle this NP-hard problem robustly. Ou r pipeline also addresses the limitation of existing 3DGS deformation algorithms and inpainting the missing information when models move around.
****

SPMTrack: Spatio-Temporal Parameter-Efficient Fine-Tuning with Mixture of Expert s for Scalable Visual Tracking

Wenrui Cai, Qingjie Liu, Yunhong Wang; Proceedings of the Computer Vision and Pa ttern Recognition Conference (CVPR), 2025, pp. 16871-16881

Most state-of-the-art trackers adopt one-stream paradigm, using a single Vision

Transformer for joint feature extraction and relation modeling of template and s
earch region images. However, relation modeling between different image patches
exhibits significant variations. For instance, background regions dominated by t
arget-irrelevant information require reduced attention allocation, while foregro
und, particularly boundary areas, need to be be emphasized. A single model may n
ot effectively handle all kinds of relation modeling simultaneously. In this pap
er, we propose a novel tracker called SPMTrack based on mixture-of-experts tailo
red for visual tracking task (TMoE), combining the capability of multiple expert
s to handle diverse relation modeling more flexibly. Benefiting from TMoE, we ex
tend relation modeling from image pairs to spatio-temporal context, further impr
oving tracking accuracy with minimal increase in model parameters. Moreover, we
employ TMoE as a parameter-efficient fine-tuning method, substantially reducing
trainable parameters, which enables us to train SPMTrack of varying scales effic
iently and preserve the generalization ability of pretrained models to achieve s
uperior performance. We conduct experiments on seven datasets, and experimental
results demonstrate that our method significantly outperforms current state-of-t
he-art trackers. The source code is available at https://github.com/WenRuiCai/SP
MTrack.
****

AnyCam: Learning to Recover Camera Poses and Intrinsics from Casual Videos
Felix Wimbauer, Weirong Chen, Dominik Muhle, Christian Rupprecht, Daniel Cremers
; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR),
2025, pp. 16717-16727
Estimating camera motion and intrinsics from casual videos is a core challenge i
n computer vision. Traditional bundle-adjustment based methods, such as SfM and
SLAM, struggle to perform reliably on arbitrary data. Although specialized SfM a
pproaches have been developed for handling dynamic scenes, they either require i
ntrinsics or computationally expensive test-time optimization and often fall sho
rt in performance. Recently, methods like Dust3r have reformulated the SfM probl
em in a more data-driven way. While such techniques show promising results, they
 are still 1) not robust towards dynamic objects and 2) require labeled data for
 supervised training. As an alternative, we propose AnyCam, a fast transformer m
odel that directly estimates camera poses and intrinsics from a dynamic video se
quence in feed-forward fashion. Our intuition is that such a network can learn s
trong priors over realistic camera poses. To scale up our training, we rely on a
n uncertainty-based loss formulation and pre-trained depth and flow networks ins
tead of motion or trajectory supervision. This allows us to use diverse, unlabel
led video datasets obtained mostly from YouTube. Additionally, we ensure that th
e predicted trajectory does not accumulate drift over time through a lightweight
 trajectory refinement step. We test AnyCam on established datasets, where it de
livers accurate camera poses and intrinsics both qualitatively and quantitativel
y. Furthermore, even with trajectory refinement, AnyCam is significantly faster
than existing works for SfM in dynamic settings. Finally, by combining camera in
formation, uncertainty, and depth, our model can produce high-quality 4D pointcl
ouds. For more details and code, please check out our project page: https://fwmb
.github.io/anycam
****

Knowledge-Aligned Counterfactual-Enhancement Diffusion Perception for Unsupervis
ed Cross-Domain Visual Emotion Recognition
Wen Yin, Yong Wang, Guiduo Duan, Dongyang Zhang, Xin Hu, Yuan-Fang Li, Tao He; P
roceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 202
5, pp. 3888-3898
Visual Emotion Recognition (VER) is a critical yet challenging task aimed at inf
erring emotional states of individuals based on visual cues. However, existing
works focus on single domains, e.g., realistic images or stickers, limiting VER
models' cross-domain generalizability. To fill this gap, we introduce an Unsupe
rvised Cross-Domain Visual Emotion Recognition (UCDVER) task, which aims to gene
ralize visual emotion recognition from the source domain (e.g., realistic images
) to the low-resource target domain (e.g., stickers) in an unsupervised manner.
Compared to the conventional unsupervised domain adaptation problems, UCDVER pre

sents two key challenges: a significant emotional expression variability and an affective distribution shift. To mitigate these issues, we propose the Knowledge-aligned Counterfactual-enhancement Diffusion Perception (KCDP) framework. Specifically, KCDP leverages a VLM to align emotional representations in a shared knowledge space and guides diffusion models for improved visual affective perception. Furthermore, a Counterfactual-Enhanced Language-image Emotional Alignment (CLIEA) method generates high-quality pseudo-labels for the target domain. Extensive experiments demonstrate that our model surpasses SOTA models in both perceptibility and generalization, e.g., gaining 12% improvements over SOTA VER model TGCA-PVT.

****

Distilling Multi-modal Large Language Models for Autonomous Driving

Deepti Hegde, Rajeev Yasarla, Hong Cai, Shizhong Han, Apratim Bhattacharyya, Shweta Mahajan, Litian Liu, Risheek Garrepalli, Vishal M. Patel, Fatih Porikli; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27575-27585

Autonomous driving demands safe motion planning, especially in critical "long-tail" scenarios. Recent end-to-end autonomous driving systems leverage large language models (LLMs) as planners to improve generalizability to rare events. However, using LLMs at test time introduces high computational costs. To address this, we propose DiMA, an end-to-end autonomous driving system that maintains the efficiency of an LLM-free (or vision-based) planner while leveraging the world knowledge of an LLM. DiMA distills the information from a multi-modal LLM to a vision-based end-to-end planner through a set of specially designed surrogate tasks. Under a joint training strategy, a scene encoder common to both networks produces structured representations that are semantically grounded as well as aligned to the final planning objective. Notably, the LLM is optional at inference, enabling robust planning without compromising on efficiency. Training with DiMA results in a 37% reduction in the L2 trajectory error and an 80% reduction in the collision rate of the vision-based planner, as well as a 44% trajectory error reduction in long-tail scenarios. \ours also achieves state-of-the-art performance on the nuScenes planning benchmark.

****

Pixel-aligned RGB-NIR Stereo Imaging and Dataset for Robot Vision

Jinnyeong Kim, Seung-Hwan Baek; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11482-11492

Integrating RGB and NIR imaging provides complementary spectral information, enhancing robotic vision in challenging lighting conditions. However, existing datasets and imaging systems lack pixel-level alignment between RGB and NIR images, posing challenges for downstream tasks.In this paper, we develop a robotic vision system equipped with two pixel-aligned RGB-NIR stereo cameras and a LiDAR sensor mounted on a mobile robot. The system simultaneously captures RGB stereo images, NIR stereo images, and temporally synchronized LiDAR point cloud. Utilizing the mobility of the robot, we present a dataset containing continuous video frames with pixel-aligned RGB and NIR stereo pairs under diverse lighting conditions.We introduce two methods that utilize our pixel-aligned RGB-NIR images: an RGB-NIR image fusion method and a feature fusion method. The first approach enables existing RGB-pretrained vision models to directly utilize RGB-NIR information without fine-tuning. The second approach fine-tunes existing vision models to more effectively utilize RGB-NIR information.Experimental results demonstrate the effectiveness of using pixel-aligned RGB-NIR images across diverse lighting conditions.

****

Can Machines Understand Composition? Dataset and Benchmark for Photographic Image Composition Embedding and Understanding

Zhaoran Zhao, Peng Lu, Anran Zhang, Peipei Li, Xia Li, Xuannan Liu, Yang Hu, Shiyi Chen, Liwei Wang, Wenhao Guo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14411-14421

With the rapid growth of social media and digital photography, visually appealing images have become essential for effective communication and emotional engagem

ent. Among the factors influencing aesthetic appeal, composition--the arrangemen t of visual elements within a frame--plays a crucial role. In recent years, spec ialized models for photographic composition have achieved impressive results acr oss various aesthetic tasks. Meanwhile, rapidly advancing multimodal large langu age models (MLLMs) have excelled in several visual perception tasks. However, th eir ability to embed and understand compositional information remains underexplo red, primarily due to the lack of suitable evaluation datasets. To address this gap, we introduce the Photographic Image Composition Dataset (PICD), a large-sca le dataset consisting of 36,857 images categorized into 24 composition categorie s across 355 diverse scenes. We demonstrate the advantages of PICD over existing datasets in terms of data scale, composition category, label quality, and scene diversity. Building on PICD, we establish benchmarks to evaluate the compositio n embedding capabilities of specialized models and the compositional understandi ng ability of MLLMs. To enable efficient and effective evaluation, we propose a novel Composition Discrimination Accuracy (CDA) metric. Our evaluation highlight s the limitations of current models and provides insights into directions for im proving their ability to embed and understand composition.

****

LPOSS: Label Propagation Over Patches and Pixels for Open-vocabulary Semantic Se gmentation

Vladan Stojni■, Yannis Kalantidis, Ji■í Matas, Giorgos Tolias; Proceedings of th e Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9794-9803

We propose a training-free method for open-vocabulary semantic segmentation usin g Vision-and-Language Models (VLMs). Our approach enhances the initial per-patch predictions of VLMs through label propagation, which jointly optimizes predicti ons by incorporating patch-to-patch relationships. Since VLMs are primarily opti mized for cross-modal alignment and not for intra-modal similarity, we use a Vis ion Model (VM) that is observed to better capture these relationships. We addres s resolution limitations inherent to patch-based encoders by applying label prop agation at the pixel level as a refinement step, significantly improving segment ation accuracy near class boundaries. Our method, called LPOSS+, performs infere nce over the entire image, avoiding window-based processing and thereby capturin g contextual interactions across the full image. LPOSS+ achieves state-of-the-ar t performance among training-free methods, across a diverse set of datasets. Cod e: https://github.com/vladan-stojnic/LPOSS

****

Towards Efficient Foundation Model for Zero-shot Amodal Segmentation

Zhaochen Liu, Limeng Qiao, Xiangxiang Chu, Lin Ma, Tingting Jiang; Proceedings o f the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20254 -20264

Aiming to predict the complete shape of partially occluded objects, amodal segme ntation is an important capacity towards visual intelligence. In order to promot e the practicability, zero-shot foundation model competent for the open world ga ins growing attention in this field. Nevertheless, prior models exhibit deficien cies in efficiency and stability. To address this problem, utilizing the implici t prior knowledge, we propose the first SAM-based amodal segmentation foundation model, SAMBA. Methodologically, a novel framework with multilevel facilitation is designed to better adapt the task characteristics and unleash the potential c apabilities of SAM. In the modality level, a separation-to-fusion structure is e mployed that jointly learns modal and amodal segmentation to enhance mutual coor dination. In the instance level, to ease the complexity of amodal feature extrac tion, we introduce a principal focusing mechanism to indicate objects of interes t. In the pixel level, mixture-of-experts is incorporated with a specialized dis tribution loss, by which distinct occlusion rates correspond to different expert s to improve the accuracy. Experiments are conducted on several eminent datasets , and the results show that the performance of SAMBA is superior to existing zer o-shot and even supervised approaches. Furthermore, our proposed model has notab le advantages in terms of speed and size.

****

PhysGen3D: Crafting a Miniature Interactive World from a Single Image

Boyuan Chen, Hanxiao Jiang, Shaowei Liu, Saurabh Gupta, Yunzhu Li, Hao Zhao, She
nlong Wang; Proceedings of the Computer Vision and Pattern Recognition Conferenc
e (CVPR), 2025, pp. 6178-6189

Envisioning physically plausible outcomes from a single image requires a deep un
derstanding of the world's dynamics. To address this, we introduce MiniTwin, a n
ovel framework that transforms a single image into an amodal, camera-centric, in
teractive 3D scene. By combining advanced image-based geometric and semantic und
erstanding with physics-based simulation, MiniTwin creates an interactive 3D wor
ld from a static image, enabling us to "imagine" and simulate future scenarios b
ased on user input. At its core, MiniTwin estimates 3D shapes, poses, physical a
nd lighting properties of objects, thereby capturing essential physical attribut
es that drive realistic object interactions. This framework allows users to spec
ify precise initial conditions, such as object speed or material properties, for
 enhanced control over generated video outcomes. We evaluate MiniTwin's performa
nce against closed-source state-of-the-art (SOTA) image-to-video models, includi
ng Pika, Kling, and Gen-3, showing MiniTwin's capacity to generate videos with r
ealistic physics while offering greater flexibility and fine-grained control. Ou
r results show that MiniTwin achieves a unique balance of photorealism, physical
 plausibility, and user-driven interactivity, opening new possibilities for gene
rating dynamic, physics-grounded video from an image.

****

Docopilot: Improving Multimodal Models for Document-Level Understanding

Yuchen Duan, Zhe Chen, Yusong Hu, Weiyun Wang, Shenglong Ye, Botian Shi, Lewei L
u, Qibin Hou, Tong Lu, Hongsheng Li, Jifeng Dai, Wenhai Wang; Proceedings of the
 Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4026-4037

Despite significant progress in multimodal large language models (MLLMs), their
 performance on complex, multi-page document comprehension remains inadequate, la
rgely due to the lack of high-quality, document-level datasets. While current re
trieval-augmented generation (RAG) methods offer partial solutions, they suffer
 from issues, such as fragmented retrieval contexts, multi-stage error accumulati
on, and extra time costs of retrieval. In this work, we present a high-quality d
ocument-level dataset, Doc-750K, designed to support in-depth understanding of m
ultimodal documents. This dataset includes diverse document structures, extensiv
e cross-page dependencies, and real question-answer pairs derived from the origi
nal documents. Building on the dataset, we develop a native multimodal model--Do
copilot, which can accurately handle document-level dependencies without relying
 on RAG. Experiments demonstrate that Docopilot achieves superior coherence, acc
uracy, and efficiency in document understanding tasks and multi-turn interaction
s, setting a new baseline for document-level multimodal understanding. Data, cod
e, and models are released at https://github.com/OpenGVLab/Docopilot.

****

Scaling Properties of Diffusion Models For Perceptual Tasks

Rahul Ravishankar, Zeeshan Patel, Jathushan Rajasegaran, Jitendra Malik; Proceed
ings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp.
 12945-12954

In this paper, we argue that iterative computation with diffusion models offers
 a powerful paradigm for not only generation but also visual perception tasks. We
 unify tasks such as depth estimation, optical flow, and amodal segmentation und
er the framework of image-to-image translation, and show how diffusion models be
nefit from scaling training and test-time compute for these perceptual tasks. Th
rough a careful analysis of these scaling properties, we formulate compute-optim
al training and inference recipes to scale diffusion models for visual perceptio
n tasks. Our models achieve competitive performance to state-of-the-art methods
using significantly less data and compute.

****

HD-EPIC: A Highly-Detailed Egocentric Video Dataset

Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti
Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, Jacob
 Chalk, Zhifan Zhu, Rhodri Guerrier, Fahd Abdelazim, Bin Zhu, Davide Moltisanti,
 Michael Wray, Hazel Doughty, Dima Damen; Proceedings of the Computer Vision and

Pattern Recognition Conference (CVPR), 2025, pp. 23901-23913
We present a validation dataset of newly-collected kitchen based egocentric videos, manually annotated with highly detailed and interconnected ground-truth labels covering: recipe steps, fine-grained actions, ingredients with nutritional values, moving objects, and audio annotations. Importantly, all annotations are grounded in 3D through digital twinning of the scene, fixtures, object locations, and primed with gaze. Footage is collected from unscripted recordings in diverse home environments, making HD-EPIC the first dataset collected in-the-wild but with detailed annotations matching those in controlled lab environments. We show the potential of our highly-detailed annotations through a challenging VQA benchmark of 26K questions assessing capability to recognise recipes, ingredients, nutrition, fine-grained actions, 3D perception, object motion, and gaze direction. The powerful long-context Gemini Pro only achieves 37.0% on this benchmark, showcasing its difficulty and highlighting shortcomings in current VLMs. We additionally assess action recognition, sound recognition, and long-term video-object segmentation on HD-EPIC. HD-EPIC is 41 hours of video in 9 kitchens with digital twins of 404 kitchen fixtures, capturing 69 recipes, 59K fine-grained actions, 51K audio events, 20K object movements and 37K object masks lifted to 3D. On average, we have 263 annotations per min of our unscripted videos.
****

Exact: Exploring Space-Time Perceptive Clues for Weakly Supervised Satellite Image Time Series Semantic Segmentation
Hao Zhu, Yan Zhu, Jiayu Xiao, Tianxiang Xiao, Yike Ma, Yucheng Zhang, Feng Dai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14036-14045
Automated crop mapping through Satellite Image Time Series (SITS) has emerged as a crucial avenue for agricultural monitoring and management. However, due to the low resolution and unclear parcel boundaries, annotating pixel-level masks is exceptionally complex and time-consuming in SITS. This paper embraces the weakly supervised paradigm (i.e., only image-level categories available) to liberate the crop mapping task from the exhaustive annotation burden. The unique characteristics of SITS give rise to several challenges in weakly supervised learning: (1) noise perturbation from spatially neighboring regions, and (2) erroneous semantic bias from anomalous temporal periods. To address the above difficulties, we propose a novel method, termed exploring space-time perceptive clues (Exact). First, we introduce a set of spatial clues to explicitly capture the representative patterns of different crops from the most class-relative regions. Besides, we leverage the temporal-to-class interaction of the model to emphasize the contributions of pivotal clips, thereby enhancing the model perception for crop regions. Build upon the space-time perceptive clues, we derive the clue-based CAMs to effectively supervise the SITS segmentation network. Our method demonstrates impressive performance on various SITS benchmarks. Remarkably, the segmentation network trained on Exact-generated masks achieves 95% of its fully supervised performance, showing the bright promise of weakly supervised paradigm in crop mapping scenario.
****

Advancing Myopia To Holism: Fully Contrastive Language-Image Pre-training
Haicheng Wang, Chen Ju, Weixiong Lin, Shuai Xiao, Mengting Chen, Yixuan Huang, Chang Liu, Mingshuai Yao, Jinsong Lan, Ying Chen, Qingwen Liu, Yanfeng Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29791-29802
In rapidly evolving field of vision-language models (VLMs), contrastive language-image pre-training (CLIP) has made significant strides, becoming foundation for various downstream tasks. However, relying on one-to-one (image, text) contrastive paradigm to learn alignment from large-scale messy web data, CLIP faces a serious myopic dilemma, resulting in biases towards monotonous short texts and shallow visual expressivity. To overcome these issues, this paper advances CLIP into one novel holistic paradigm, by updating both diverse data and alignment optimization. To obtain colorful data with low cost, we use image-to-text captioning to generate multi-texts for each image, from multiple perspectives, granularitie

s, and hierarchies. Two gadgets are proposed to encourage textual diversity. To match such (image, multi-texts) pairs, we modify the CLIP image encoder into multi-branch, and propose multi-to-multi contrastive optimization for image-text part-to-part matching. As a result, diverse visual embeddings are learned for each image, bringing good interpretability and generalization. Extensive experiments and ablations across over ten benchmarks indicate that our holistic CLIP significantly outperforms existing myopic CLIP, including image-text retrieval, open-vocabulary classification, and dense visual tasks. Project page is available to further promote the prosperity of VLMs: https://voide1220.github.io/Holism/.
****

PolarFree: Polarization-based Reflection-Free Imaging
Mingde Yao, Menglu Wang, King-Man Tam, Lingen Li, Tianfan Xue, Jinwei Gu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10890-10899
Reflection removal is challenging due to complex light interactions, where reflections obscure important details and hinder scene understanding. Polarization naturally provides a powerful cue to distinguish between reflected and transmitted light, enabling more accurate reflection removal. However, existing methods often rely on small-scale or synthetic datasets, which fail to capture the diversity and complexity of real-world scenarios. To this end, we construct a large-scale dataset, PolaRGB, for Polarization-based reflection removal of RGB images, which enables us to train models that generalize effectively across a wide range of real-world scenarios. The PolaRGB dataset contains 6,500 well-aligned mixed-transmission image pairs, 8xlarger than existing polarization datasets, and is the first to include both RGB and polarization images captured across diverse indoor and outdoor environments with varying lighting conditions. Besides, to fully exploit the potential of polarization cues for reflection removal, we introduce PolarFree, which leverages diffusion process to generate reflection-free cues for accurate reflection removal. Extensive experiments show that PolarFree significantly enhances image clarity in challenging reflective scenarios, setting a new benchmark for polarized imaging and reflection removal. Code and dataset are available at https://github.com/mdyao/PolarFree.
****

H-MoRe: Learning Human-centric Motion Representation for Action Analysis
Zhanbo Huang, Xiaoming Liu, Yu Kong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22702-22713
In this paper, we propose H-MoRe, a novel pipeline for learning precise human-centric motion representation. Our approach dynamically preserves relevant human motion while filtering out background movement. Notably, unlike previous methods relying on fully supervised learning from synthetic data, H-MoRe learns directly from real-world scenarios in a self-supervised manner, incorporating both human pose and body shape information. Inspired by kinematics, H-MoRe represents absolute and relative movements of each body point in a matrix format that captures nuanced motion details, termed world-local flows. H-MoRe offers refined insights into human motion, which can be integrated seamlessly into various action-related applications. Experimental results demonstrate that H-MoRe brings substantial improvements across various downstream tasks, including gait recognition(CL@R1: +16.01%), action recognition(Acc@1: +8.92%), and video generation(FVD: -67.07%). Additionally, H-MoRe exhibits high inference efficiency (34 fps), making it suitable for most real-time scenarios. Models and code will be released upon publication.
****

Hierarchical Compact Clustering Attention (COCA) for Unsupervised Object-Centric Learning
Can Kucuksozen, Yucel Yemez; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25388-25398
We propose the Compact Clustering Attention (COCA) layer, an effective building block that introduces a hierarchical strategy for object-centric representation learning, while solving the unsupervised object discovery task on single images. COCA is an attention-based clustering module capable of extracting object-centr

ic representations from multi-object scenes, when cascaded into a bottom-up hier archical network architecture, referred to as COCA-Net. At its core, COCA utiliz es a novel clustering algorithm that leverages the physical concept of compactne ss, to highlight distinct object centroids in a scene, providing a spatial induc tive bias. Thanks to this strategy, COCA-Net generates high-quality segmentation masks on both the decoder side and, notably, the encoder side of its pipeline. Additionally, COCA-Net is not bound by a predetermined number of object masks th at it generates and handles the segmentation of background elements better than its competitors. We demonstrate COCA-Net's segmentation performance on six widel y adopted datasets, achieving superior or competitive results against the state-of-the-art models across nine different evaluation metrics.

****

Effortless Active Labeling for Long-Term Test-Time Adaptation

Guowei Wang, Changxing Ding; Proceedings of the Computer Vision and Pattern Reco gnition Conference (CVPR), 2025, pp. 25633-25642

Long-term test-time adaptation (TTA) is a challenging task due to error accumula tion. Recent approaches tackle this issue by actively labeling a small proportio n of samples in each batch, yet the annotation burden quickly grows as the batch number increases. In this paper, we investigate how to achieve effortless activ e labeling so that a maximum of one sample is selected for annotation in each ba tch. First, we annotate the most valuable sample in each batch based on the sing le-step optimization perspective in the TTA context. In this scenario, the sampl es that border between the source- and target-domain data distributions are cons idered the most feasible for the model to learn in one iteration. Then, we intro duce an efficient strategy to identify these samples using feature perturbation. Second, we discover that the gradient magnitudes produced by the annotated and unannotated samples have significant variations. Therefore, we propose balancing their impact on model optimization using two dynamic weights. Extensive experim ents on the popular ImageNet-C, -R, -K, -A and PACS databases demonstrate that o ur approach consistently outperforms state-of-the-art methods with significantly lower annotation costs. Code is available at: \href https://github.com/flash18 03/EATTA https://github.com/flash1803/EATTA .

****

Leveraging Temporal Cues for Semi-Supervised Multi-View 3D Object Detection

Jinhyung Park, Navyata Sanghvi, Hiroki Adachi, Yoshihisa Shibata, Shawn Hunt, Sh inya Tanaka, Hironobu Fujiyoshi, Kris Kitani; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27401-27412

While recent advancements in camera-based 3D object detection demonstrate remark able performance, they require thousands or even millions of human-annotated fra mes. This requirement significantly inhibits their deployment in various locatio ns and sensor configurations. To address this gap, we propose a performant semi-supervised framework that leverages unlabeled RGB-only driving sequences - data easily collected with cost-effective RGB cameras - to significantly improve temp oral, camera-only 3D detectors. We observe that the standard semi-supervised pse udo-labeling paradigm underperforms in this temporal, camera-only setting due to poor 3D localization of pseudo-labels. To address this, we train a single 3D de tector to handle RGB sequences both forward and backward in time, then ensemble both its forwards and backwards pseudo-labels for semi-supervised learning. We f urther improve the pseudo-label quality by leveraging 3D object tracking to infi ll missing detections and by eschewing simple confidence thresholding in favor o f using the auxiliary 2D detection head to filter 3D predictions. Finally, to en able the backbone to learn directly from the unlabeled data itself, we introduce an object-query conditioned masked reconstruction objective. Our framework demo nstrates remarkable performance improvement on large-scale autonomous driving da tasets nuScenes and nuPlan.

****

Self-supervised ControlNet with Spatio-Temporal Mamba for Real-world Video Super -resolution

Shijun Shi, Jing Xu, Lijing Lu, Zhihang Li, Kai Hu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7385-7395

Existing diffusion-based video super-resolution (VSR) methods are susceptible to introducing complex degradations and noticeable artifacts into high-resolution videos due to their inherent randomness. In this paper, we propose a noise-robust real-world VSR framework by incorporating self-supervised learning and Mamba into pre-trained latent diffusion models. To ensure content consistency across adjacent frames, we enhance the diffusion model with a global spatio-temporal attention mechanism using the Video State-Space block with a 3D Selective Scan module, which reinforces coherence at an affordable computational cost. To further reduce artifacts in generated details, we introduce a self-supervised ControlNet that leverages HR features as guidance and employs contrastive learning to extract degradation-insensitive features from LR videos. Finally, a three-stage training strategy based on a mixture of HR-LR videos is proposed to stabilize VSR training. The proposed Self-supervised ControlNet with Spatio-Temporal Continuous Mamba based VSR algorithm achieves superior perceptual quality than state-of-the-arts on real-world VSR benchmark datasets, validating the effectiveness of the proposed model design and training strategies.

****

LATTE-MV: Learning to Anticipate Table Tennis Hits from Monocular Videos

Daniel Etaat, Dvij Kalaria, Nima Rahmanian, S. Shankar Sastry; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7115-7124

Physical agility is a necessary skill in competitive table tennis, but by no means sufficient. Champions excel in this fast-paced and highly dynamic environment by anticipating their opponent's intent - buying themselves the necessary time to react. In this work, we take one step towards designing such an anticipatory agent. Previous works have developed systems capable of real-time table tennis gameplay, though they often do not leverage anticipation. Among the works that forecast opponent actions, their approaches are limited by dataset size and variety. Our paper contributes (1) a scalable system for reconstructing monocular video of table tennis matches in 3D and (2) an uncertainty-aware controller that anticipates opponent actions. We demonstrate in simulation that our policy improves the ball return rate against high-speed hits from 49.9% to 59.0% as compared to a baseline non-anticipatory policy.

****

Logits DeConfusion with CLIP for Few-Shot Learning

Shuo Li, Fang Liu, Zehua Hao, Xinyi Wang, Lingling Li, Xu Liu, Puhua Chen, Wenping Ma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25411-25421

With its powerful visual-language alignment capability, CLIP performs well in zero-shot and few-shot learning tasks. However, we found in experiments that CLIP's logits suffer from serious inter-class confusion problems in downstream tasks, and the ambiguity between categories seriously affects the accuracy. To address this challenge, we propose a novel method called Logits DeConfusion, which effectively learns and eliminates inter-class confusion in logits by combining our Multi-level Adapter Fusion (MAF) module with our Inter-Class Deconfusion (ICD) module. Our MAF extracts features from different levels and fuses them uniformly to enhance feature representation. Our ICD learnably eliminates inter-class confusion in logits with a residual structure. Experimental results show that our method can significantly improve the classification performance and alleviate the inter-class confusion problem. The code is available at https://github.com/LiShuo1001/LDC.

****

Distilling Spatially-Heterogeneous Distortion Perception for Blind Image Quality Assessment

Xudong Li, Wenjie Nie, Yan Zhang, Runze Hu, Ke Li, Xiawu Zheng, Liujuan Cao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2344-2354

In the Blind Image Quality Assessment (BIQA) field, accurately assessing the quality of authentically distorted images presents a substantial challenge due to the diverse distortion types in natural settings. Existing state-of-the-art IQA methods mix a sequence of distortions into entire images to establish global dist

ortion priors, but are inadequate for authentic images with spatially varied dis tortions. To address this, we introduce a novel IQA framework that employs knowl edge distillation tailored to perceive spatially heterogeneous distortions, enha ncing quality-distortion awareness. Specifically, we introduce a novel Block-wis e Degradation Modelling approach that applies distinct distortions to different spatial blocks of an image, thereby expanding local distortion priors. Following this, we present a Block-wise Aggregation and Filtering module that enables fin e-grained attention to the quality information within different distortion areas of the image. Furthermore, to effectively capture the complex relationships bet ween distortions across different regions while preserving overall quality perce ption, we introduce Contrastive Knowledge Distillation to enhance the model's ab ility to discriminate between different types of distortions and Affinity Knowle dge Distillation to model the correlation among distortions in different regions . Extensive experiments on standard BIQA datasets demonstrate the effectiveness and competitiveness of the proposed method.

****

Pay Attention to the Foreground in Object-Centric Learning

Pinzhuo Tian, Shengjie Yang, Hang Yu, Alex Kot; Proceedings of the Computer Visi on and Pattern Recognition Conference (CVPR), 2025, pp. 30281-30290

The slot attention-based method is widely used in unsupervised object-centric le arning, aiming to decompose scenes into interpretable objects and associate them with slots. However, complex backgrounds in the real images can disrupt the mod el's focus, leading it to excessively segment background stuff into different re gions based on low-level information such as color or texture variations. As a r esult, the detailed segmentation of foreground objects, which requires shape or geometric information, is often neglected. To address this issue, we introduce a contrastive learning-based indicator designed to differentiate between foregro und and background. Integrating this indicator into the existing slot attention- based method enables the model to focus more on segmenting foreground objects wh ile minimizing background distractions. During the testing phase, we utilize a s pectral clustering mechanism to refine the results based on the similarity betwe en the slots. Experimental results show that incorporating our method with vario us state-of-the-art models significantly improves their performance on both simu lated data and real-world datasets. Furthermore, multiple sets of ablation exper iments confirm the effectiveness of each proposed component.

****

2DMamba: Efficient State Space Model for Image Representation with Applications on Giga-Pixel Whole Slide Image Classification

Jingwei Zhang, Anh Tien Nguyen, Xi Han, Vincent Quoc-Huy Trinh, Hong Qin, Dimitr is Samaras, Mahdi S. Hosseini; Proceedings of the Computer Vision and Pattern Re cognition Conference (CVPR), 2025, pp. 3583-3592

Efficiently modeling large 2D contexts is essential for various fields including Giga-Pixel Whole Slide Imaging (WSI) and remote sensing. Transformer-based mode ls offer high parallelism but face challenges due to their quadratic complexity for handling long sequences. Recently, Mamba introduced a selective State Space Model (SSM) with linear complexity and high parallelism, enabling effective and efficient modeling of wide context in 1D sequences. However, extending Mamba to vision tasks, which inherently involve 2D structures, results in spatial discrep ancies due to the limitations of 1D sequence processing. On the other hand, curr ent 2D SSMs inherently model 2D structures but they suffer from prohibitively sl ow computation due to the lack of efficient parallel algorithms. In this work, w e propose 2DMamba, a novel 2D selective SSM framework that incorporates the 2D s patial structure of images into Mamba, with a highly optimized hardware-aware op erator, adopting both spatial continuity and computational efficiency. We valida te the versatility of our approach on both WSIs and natural images. Extensive ex periments on 10 public datasets for WSI classification and survival analysis sho w that 2DMamba improves up to 2.48% in AUC, 3.11% in F1 score, 2.47% in accuracy and 5.52% in C-index. Additionally, integrating our method with VMamba for natu ral imaging yields 0.5 to 0.7 improvements in mIoU on the ADE20k semantic segmen tation dataset, and 0.2% accuracy improvement on ImageNet-1K classification data

set. Our code is available at https://github.com/AtlasAnalyticsLab/2DMamba.
****
Unboxed: Geometrically and Temporally Consistent Video Outpainting
Zhongrui Yu, Martina Megaro-Boldini, Robert W. Sumner, Abdelaziz Djelouah; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7309-7319
Extending the field of view of video content beyond its original version has many applications: immersive viewing experience with VR devices, reformatting 4:3 legacy content to today's viewing conditions with wide screens, or simply extending vertically captured phone videos. Many existing works focus on synthesizing the video using generative models only. Despite promising results, this strategy seems at the moment limited in terms of quality. In this work, we address this problem using two key ideas: 3D supported outpainting for the static regions of the images, and leveraging pre-trained video diffusion model to ensure realistic and temporally coherent results, particularly for the dynamic parts. In the first stage, we iterate between image outpainting and updating the 3D scene representation - we use 3D Gaussian Splatting. Then we consider dynamic objects independently per frame and inpaint missing pixels. Finally, we propose a denoising scheme that allows to maintain known reliable regions and update the dynamic parts to obtain temporally realistic results.We achieve state-of-the-art video outpainting. This is validated quantitatively and through a user study. We are also able to extend the field of view largely beyond the limits reached by existing methods.
****
K-Sort Arena: Efficient and Reliable Benchmarking for Generative Models via K-wise Human Preferences
Zhikai Li, Xuewen Liu, Dongrong Joe Fu, Jianquan Li, Qingyi Gu, Kurt Keutzer, Zhen Dong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9131-9141
The rapid advancement of visual generative models necessitates efficient and reliable evaluation methods. Arena platform, which gathers user votes on model comparisons, can rank models with human preferences. However, traditional Arena methods, while established, require an excessive number of comparisons for ranking to converge and are vulnerable to preference noise in voting, suggesting the need for better approaches tailored to contemporary evaluation challenges. In this paper, we introduce K-Sort Arena, an efficient and reliable platform based on a key insight: images and videos possess higher perceptual intuitiveness than texts, enabling rapid evaluation of multiple samples simultaneously. Consequently, K-Sort Arena employs K-wise comparisons, allowing K models to engage in free-for-all competitions, which yield much richer information than pairwise comparisons. To enhance the robustness of the system, we leverage probabilistic modeling and Bayesian updating techniques. We propose an exploration-exploitation-based matchmaking strategy to facilitate more informative comparisons. In our experiments, K-Sort Arena exhibits 16.3xfaster convergence compared to the widely used ELO algorithm. To further validate the superiority and obtain a comprehensive leaderboard, we collect human feedback via crowdsourced evaluations of numerous cutting-edge text-to-image and text-to-video models. Thanks to its high efficiency, K-Sort Arena can continuously incorporate emerging models and update the leaderboard with minimal votes. Our project has undergone several months of internal testing and is now available online.
****
Seeking Consistent Flat Minima for Better Domain Generalization via Refining Loss Landscapes
Aodi Li, Liansheng Zhuang, Xiao Long, Minghong Yao, Shafei Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15349-15359
Domain generalization aims to learn a model from multiple training domains and generalize it to unseen test domains. Recent theory has shown that seeking the deep models, whose parameters lie in the flat minima of the loss landscape, can significantly reduce the out-of-domain generalization error. However, existing met

hods often neglect the consistency of loss landscapes in different domains, resu
lting in models that are not simultaneously in the optimal flat minima in all do
mains, which limits their generalization ability. To address this issue, this pa
per proposes an iterative Self-Feedback Training (SFT) framework to seek consist
ent flat minima that are shared across different domains by progressively refini
ng loss landscapes during training. It alternatively generates a feedback signal
 by measuring the inconsistency of loss landscapes in different domains and refi
nes these loss landscapes for greater consistency using this feedback signal. Be
nefiting from the consistency of the flat minima within these refined loss lands
capes, our SFT helps achieve better out-of-domain generalization. Extensive expe
riments on DomainBed demonstrate superior performances of SFT when compared to s
tate-of-the-art sharpness-aware methods and other prevalent DG baselines. On ave
rage across five DG benchmarks, SFT surpasses the sharpness-aware minimization b
y 2.6% with ResNet-50 and 1.5% with ViT-B/16, respectively.
****

MultimodalStudio: A Heterogeneous Sensor Dataset and Framework for Neural Render
ing across Multiple Imaging Modalities
Federico Lincetto, Gianluca Agresti, Mattia Rossi, Pietro Zanuttigh; Proceedings
 of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 109
64-10973
Neural Radiance Fields (NeRF) have shown impressive performances in the renderin
g of 3D scenes from arbitrary viewpoints. While RGB images are widely preferred
for training volume rendering models, the interest in other radiance modalities
is also growing. However, the capability of the underlying implicit neural model
s to learn and transfer information across heterogeneous imaging modalities has
seldom been explored, mostly due to the limited training data availability. For
this purpose, we present MultimodalStudio (MMS): it encompasses MMS-DATA and MMS
-FW. MMS-DATA is a multimodal multi-view dataset containing 32 scenes acquired w
ith 5 different imaging modalities: RGB, monochrome, near-infrared, polarization
 and multispectral. MMS-FW is a novel modular multimodal NeRF framework designed
 to handle multimodal raw data and able to support an arbitrary number of multi-
channel devices. Through extensive experiments, we demonstrate that MMS-FW train
ed on MMS-DATA can transfer information between different imaging modalities and
 produce higher quality renderings than using single modalities alone. We public
ly release the dataset and the framework, to promote the research on multimodal
volume rendering and beyond.
****

Dense-SfM: Structure from Motion with Dense Consistent Matching
JongMin Lee, Sungjoo Yoo; Proceedings of the Computer Vision and Pattern Recogni
tion Conference (CVPR), 2025, pp. 6404-6414
We present Dense-SfM, a novel Structure from Motion (SfM) framework designed for
 dense and accurate 3D reconstruction from multi-view images. Sparse keypoint ma
tching, which traditional SfM methods often rely on, limits both accuracy and po
int density, especially in texture-less areas. Dense-SfM addresses this limitati
on by integrating dense matching with a Gaussian Splatting (GS) based track exte
nsion which gives more consistent, longer feature tracks. To further improve rec
onstruction accuracy, Dense-SfM is equipped with a multi-view kernelized matchin
g module leveraging transformer and Gaussian Process architectures, for robust t
rack refinement across multi-views. Evaluations on the ETH3D and Texture-Poor Sf
M datasets show that Dense-SfM offers significant improvements in accuracy and d
ensity over state-of-the-art methods. Project page: https://icetea-cv.github.io/
densesfm/.
****

FluidNexus: 3D Fluid Reconstruction and Prediction from a Single Video
Yue Gao, Hong-Xing Yu, Bo Zhu, Jiajun Wu; Proceedings of the Computer Vision and
 Pattern Recognition Conference (CVPR), 2025, pp. 26091-26101
We study reconstructing and predicting 3D fluid appearance and velocity from a s
ingle video. Current methods require multi-view videos for fluid reconstruction.
 We present FluidNexus, a novel framework that bridges video generation and phys
ics simulation to tackle this task. Our key insight is to synthesize multiple no

vel-view videos as references for reconstruction. FluidNexus consists of two key components: (1) a novel-view video synthesizer that combines frame-wise view synthesis with video diffusion refinement for generating realistic videos, and (2) a physics-integrated particle representation coupling differentiable simulation and rendering to simultaneously facilitate 3D fluid reconstruction and prediction. To evaluate our approach, we collect two new real-world fluid datasets featuring textured backgrounds and object interactions. Our method enables dynamic novel view synthesis, future prediction, and interaction simulation from a single fluid video. Project website: https://yuegao.me/FluidNexus.
****

MuTri: Multi-view Tri-alignment for OCT to OCTA 3D Image Translation
Zhuangzhuang Chen, Hualiang Wang, Chubin Ou, Xiaomeng Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20885-20894
Optical coherence tomography angiography (OCTA) shows its great importance in imaging microvascular networks by providing accurate 3D imaging of blood vessels, but it relies upon specialized sensors and expensive devices. For this reason, previous works show the potential to translate the readily available 3D Optical Coherence Tomography (OCT) images into 3D OCTA images. However, existing OCTA translation methods directly learn the mapping from the OCT domain to the OCTA domain in continuous and infinite space with guidance from only a single view, i.e., the OCTA project map, resulting in suboptimal results. To this end, we propose the multi-view Tri-alignment framework for OCT to OCTA 3D image translation in discrete and finite space, named MuTri. In the first stage, we pre-train two vector-quantized variational auto-encoder (VQVAE) by reconstructing 3D OCT and 3D OCTA data, providing semantic prior for subsequent multi-view guidances. In the second stage, our multi-view tri-alignment facilitates another VQVAE model to learn the mapping from the OCT domain to the OCTA domain in discrete and finite space. Specifically, a contrastive-inspired semantic alignment is proposed to maximize the mutual information with the pre-trained models from OCT and OCTA views, to facilitate codebook learning. Meanwhile, a vessel structure alignment is proposed to minimize the structure discrepancy with the pre-trained models from the OCTA project map view, benefiting from learning the detailed vessel structure information. We also collect the first large-scale dataset, namely, OCTA2024, which contains a pair of OCT and OCTA volumes from 846 subjects. Our codes and datasets are available at: https://github.com/xmed-lab/MuTri.
****

Sketchy Bounding-box Supervision for 3D Instance Segmentation
Qian Deng, Le Hui, Jin Xie, Jian Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8879-8888
Bounding box supervision has gained considerable attention in weakly supervised 3D instance segmentation. While this approach alleviates the need for extensive point-level annotations, obtaining accurate bounding boxes in practical applications remains challenging. To this end, we explore the inaccurate bounding box, named sketchy bounding box, which is imitated through perturbing ground truth bounding box by adding scaling, translation, and rotation. In this paper, we propose Sketchy-3DIS, a novel weakly 3D instance segmentation framework, which jointly learns pseudo labeler and segmentator to improve the performance under the sketchy bounding-box supervisions. Specifically, we first propose an adaptive box-to-point pseudo labeler that adaptively learns to assign points located in the overlapped parts between two sketchy bounding boxes to the correct instance, resulting in compact and pure pseudo instance labels. Then, we present a coarse-to-fine instance segmentator that first predicts coarse instances from the entire point cloud and then learns fine instances based on the region of coarse instances. Finally, by using the pseudo instance labels to supervise the instance segmentator, we can gradually generate high-quality instances through joint training. Extensive experiments show that our method achieves state-of-the-art performance on both the ScanNetV2 and S3DIS benchmarks, and even outperforms several fully supervised methods using sketchy bounding boxes. Code is available at https://github.com/dengq7/Sketchy-3DIS.
****

Image Quality Assessment: Investigating Causal Perceptual Effects with Abductive Counterfactual Inference

Wenhao Shen, Mingliang Zhou, Yu Chen, Xuekai Wei, Yong Feng, Huayan Pu, Weijia Jia; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17990-17999

Existing full-reference image quality assessment (FR-IQA) methods often fail to capture the complex causal mechanisms that underlie human perceptual responses to image distortions, limiting their ability to generalize across diverse scenarios. In this paper, we propose an FR-IQA method based on abductive counterfactual inference to investigate the causal relationships between deep network features and perceptual distortions. First, we explore the causal effects of deep features on perception and integrate causal reasoning with feature comparison, constructing a model that effectively handles complex distortion types across different IQA scenarios. Second, the analysis of the perceptual causal correlations of our proposed method is independent of the backbone architecture and thus can be applied to a variety of deep networks. Through abductive counterfactual experiments, we validate the proposed causal relationships, confirming the model's superior perceptual relevance and interpretability of quality scores. The experimental results demonstrate the robustness and effectiveness of the method, providing competitive quality predictions across multiple benchmarks. The source code is available at https://anonymous.4open.science/r/DeepCausalQuality-25BC.

****

Pos3R: 6D Pose Estimation for Unseen Objects Made Easy

Weijian Deng, Dylan Campbell, Chunyi Sun, Jiahao Zhang, Shubham Kanitkar, Matt E. Shaffer, Stephen Gould; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16818-16828

Foundation models have significantly reduced the need for task-specific training, while also enhancing generalizability. However, state-of-the-art 6D pose estimators either require further training with pose supervision or neglect advances obtainable from 3D foundation models. The latter is a missed opportunity, since these models are better equipped to predict 3D-consistent features, which are of significant utility for the pose estimation task. To address this gap, we propose Pos3R, a method for estimating the 6D pose of any object from a single RGB image, making extensive use of a 3D reconstruction foundation model and requiring no additional training. We identify template selection as a particular bottleneck for existing methods that is significantly alleviated by the use of a 3D model, which can more easily distinguish between template poses than a 2D model. Despite its simplicity, Pos3R achieves competitive performance on the Benchmark for 6D Object Pose Estimation (BOP), matching or surpassing existing refinement-free methods. Additionally, Pos3R integrates seamlessly with render-and-compare refinement techniques, demonstrating adaptability for high-precision applications.

****

DeformCL: Learning Deformable Centerline Representation for Vessel Extraction in 3D Medical Image

Ziwei Zhao, Zhixing Zhang, Yuhang Liu, Zhao Zhang, Haojun Yu, Dong Wang, Liwei Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30896-30905

In the field of 3D medical imaging, accurately extracting and representing the blood vessels with curvilinear structures holds paramount importance for clinical diagnosis. Previous methods have commonly relied on discrete representation like mask, often resulting in local fractures or scattered fragments due to the inherent limitations of the per-pixel classification paradigm. In this work, we introduce DeformCL, a new continuous representation based on Deformable Centerlines, where centerline points act as nodes connected by edges that capture spatial relationships. Compared with previous representations, DeformCL offers three key advantages: natural connectivity, noise robustness, and interaction facility. We present a comprehensive training pipeline structured in a cascaded manner to fully exploit these favorable properties of DeformCL. Extensive experiments on four 3D vessel segmentation datasets demonstrate the effectiveness and superiority of our method. Furthermore, the visualization of curved planar reformation image

s validates the clinical significance of the proposed framework.
****

StreetCrafter: Street View Synthesis with Controllable Video Diffusion Models
Yunzhi Yan, Zhen Xu, Haotong Lin, Haian Jin, Haoyu Guo, Yida Wang, Kun Zhan, Xia npeng Lang, Hujun Bao, Xiaowei Zhou, Sida Peng; Proceedings of the Computer Visi on and Pattern Recognition Conference (CVPR), 2025, pp. 822-832
This paper aims to tackle the problem of photorealistic view synthesis from vehi cle sensors data. Recent advancements in neural scene representation have achiev ed notable success in rendering high-quality autonomous driving scenes,but the p erformance significantly degrades as the viewpoint deviates from the training tr ajectory. To mitigate this problem, we introduce StreetCrafter, a novel controll able video diffusion model that utilizes LiDAR point cloud renderings as pixel-l evel conditions, which fully exploits the generative prior for novel view synthe sis, while preserving precise camera control.Moreover, the utilization of pixel- level LiDAR condition allows us to make accurate pixel-level edits to target sce nes.In addition, the generative prior of StreetCrafter can be effectively incorp orated into dynamic scene representations to achieve real-time rendering. Experi ments on Waymo Open and Pandaset datasets demonstrate that our model enables fle xible control over viewpoint changes, enlarging the view synthesis regions for s atisfying rendering, which outperforms existing methods.
****

OCRT: Boosting Foundation Models in the Open World with Object-Concept-Relation Triad
Luyao Tang, Yuxuan Yuan, Chaoqi Chen, Zeyu Zhang, Yue Huang, Kun Zhang; Proceedi ngs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25422-25433
Although foundation models (FMs) claim to be powerful, their generalization abil ity significantly decreases when faced with distribution shifts, weak supervisio n, or malicious attacks in the open world. On the other hand, most domain genera lization or adversarial fine-tuning methods are task-related or model-specific, ignoring the universality in practical applications and the transferability betw een FMs. This paper delves into the problem of generalizing FMs to the out-of-do main data. We propose a novel framework, Object-Concept-Relation Triad (OCRT), t hat enables FMs to extract sparse, high-level concepts and intricate relational structures from raw visual inputs. The key idea is to bind objects in visual sce nes and a set of object-centric representations through unsupervised decoupling and iterative refinement. To be specific, we project the object-centric represen tations onto a semantic concept space that the model can readily interpret, and estimate their importance to filter out irrelevant elements. Then, a concept-bas ed graph, which has a flexible degree, is constructed to incorporate the set of concepts and their corresponding importance, enabling the extraction of high-ord er factors from informative concepts and facilitating relational reasoning among these concepts. Extensive experiments demonstrate that OCRT can substantially b oost the generalizability and robustness of SAM and CLIP across multiple downstr eam tasks.
****

SPARS3R: Semantic Prior Alignment and Regularization for Sparse 3D Reconstructio n
Yutao Tang, Yuxiang Guo, Deming Li, Cheng Peng; Proceedings of the Computer Visi on and Pattern Recognition Conference (CVPR), 2025, pp. 26810-26821
Recent efforts in Gaussian-Splat-based Novel View Synthesis can achieve photorea listic rendering; however, such capability is limited in sparse-view scenarios d ue to sparse initialization and over-fitting floaters. Recent progress in depth estimation and alignment can provide dense point cloud using few views; however, the resulting pose accuracy is suboptimal. In this work, we present SPARS3R, wh ich combines the advantages of accurate pose estimation from Structure-from-Moti on and dense point cloud from depth estimation. To this end, SPARS3R first perfo rms a Global Fusion Alignment process that maps a prior dense point cloud to a s parse point cloud from Structure-from-Motion based on triangulated correspondenc es. RANSAC is applied during this process to distinguish inliers and outliers. S

PARS3R then performs a second, Semantic Outlier Alignment step, which extracts s
emantically coherent regions around the outliers and performs local alignment in
 these regions. Along with several improvements in the evaluation process, we de
monstrate that SPARS3R can achieve photorealistic rendering with sparse images a
nd significantly outperforms existing approaches.
****

VidBot: Learning Generalizable 3D Actions from In-the-Wild 2D Human Videos for Z
ero-Shot Robotic Manipulation
Hanzhi Chen, Boyang Sun, Anran Zhang, Marc Pollefeys, Stefan Leutenegger; Procee
dings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp
. 27661-27672
Future robots are envisioned as versatile systems capable of performing a variet
y of household tasks. The big question remains, how can we bridge the embodiment
 gap while minimizing physical robot learning, which fundamentally does not scal
e well. We argue that learning from in-the-wild human videos offers a promising
solution for robotic manipulation tasks, as vast amounts of relevant data alread
y exist on the internet. In this work, we present VidBot, a framework enabling z
ero-shot robotic manipulation using learned 3D affordance from in-the-wild monoc
ular RGB-only human videos. VidBot leverages a pipeline to extract explicit repr
esentations from them, namely 3D hand trajectories from videos, combining a dept
h foundation model with structure-from-motion techniques to reconstruct temporal
ly consistent, metric-scale 3D affordance representations agnostic to embodiment
s. We introduce a coarse-to-fine affordance learning model that first identifies
 coarse actions from the pixel space and then generates fine-grained interaction
 trajectories with a diffusion model, conditioned on coarse actions and guided b
y test-time constraints for context-aware interaction planning, enabling substan
tial generalization to novel scenes and embodiments. Extensive experiments demon
strate the efficacy of VidBot, which significantly outperforms counterparts acro
ss 13 manipulation tasks in zero-shot settings and can be seamlessly deployed ac
ross robot systems in real-world environments. VidBot paves the way for leveragi
ng everyday human videos to make robot learning more scalable.
****

Learning Person-Specific Animatable Face Models from In-the-Wild Images via a Sh
ared Base Model
Yuxiang Mao, Zhenfeng Fan, ZhiJie Zhang, Zhiheng Zhang, Shihong Xia; Proceedings
 of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 560
2-5613
Training a generic 3D face reconstruction model in a self-supervised manner usin
g large-scale, in-the-wild 2D face image datasets enhances robustness to varying
 lighting conditions and occlusions while allowing the model to capture animatab
le wrinkle details across diverse facial expressions. However, a generic model o
ften fails to adequately represent the unique characteristics of specific indivi
duals. In this paper, we propose a method to train a generic base model and then
 transfer it to yield person-specific models by integrating lightweight adapters
 within the large-parameter ViT-MAE base model. These person-specific models exc
el at capturing individual facial shapes and detailed features while preserving
the robustness and prior knowledge of detail variations from the base model. Dur
ing training, we introduce a silhouette vertex re-projection loss to address bou
ndary "landmark marching" issues on the 3D face caused by pose variations. Addit
ionally, we employ an innovative teacher-student loss to leverage the inherent s
trengths of UNet in feature boundary localization for training our detail MAE. Q
uantitative and qualitative experiments demonstrate that our approach achieves s
tate-of-the-art performance in face alignment, detail accuracy, and richness. Th
e source code is available at https://github.com/danielmao2000/person-specific-a
nimatable-face.
****

TIMotion: Temporal and Interactive Framework for Efficient Human-Human Motion Ge
neration
Yabiao Wang, Shuo Wang, Jiangning Zhang, Ke Fan, Jiafu Wu, Zhucun Xue, Yong Liu;
 Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2

025, pp. 7169-7178
Human-human motion generation is essential for understanding humans as social beings. Current methods fall into two main categories: single-person-based methods and separate modeling-based methods. To delve into this field, we abstract the overall generation process into a general framework MetaMotion, which consists of two phases: temporal modeling and interaction mixing. For temporal modeling, the single-person-based methods concatenate two people into a single one directly, while the separate modeling-based methods skip the modeling of interaction sequences. The inadequate modeling described above resulted in sub-optimal performance and redundant model parameters. In this paper, we introduce TIMotion (Temporal and Interactive Modeling), an efficient and effective framework for human-human motion generation. Specifically, we first propose Causal Interactive Injection to model two separate sequences as a causal sequence leveraging the temporal and causal properties. Then we present Role-Evolving Scanning to adjust to the change in the active and passive roles throughout the interaction. Finally, to generate smoother and more rational motion, we design Localized Pattern Amplification to capture short-term motion patterns.Extensive experiments on InterHuman and InterX demonstrate that our method achieves superior performance.
****
Which Viewpoint Shows it Best? Language for Weakly Supervising View Selection in Multi-view Instructional Videos
Sagnik Majumder, Tushar Nagarajan, Ziad Al-Halah, Reina Pradhan, Kristen Grauman; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29016-29028
Given a multi-view video, which viewpoint is most informative for a human observer? Existing methods rely on heuristics or expensive "best-view" supervision to answer this question, limiting their applicability. We propose a weakly supervised approach that leverages language accompanying an instructional multi-view video as a means to recover its most informative viewpoint(s). Our key hypothesis is that the more accurately an individual view can predict a view-agnostic text summary, the more informative it is. To put this into action, we propose LangView, a framework that uses the relative accuracy of view dependent caption predictions as a proxy for best view pseudo-labels. Then, those pseudo-labels are used to train a view selector, together with an auxiliary camera pose predictor that enhances view-sensitivity. During inference, our model takes as input only a multi-view video--no language or camera poses--and returns the best viewpoint to watch at each timestep. On two challenging datasets comprised of diverse multi-camera setups and how-to activities, our model consistently outperforms state-of-the-art baselines, both with quantitative metrics and human evaluation. Project: https://vision.cs.utexas.edu/projects/which-view-shows-it-best.
****
RaCFormer: Towards High-Quality 3D Object Detection via Query-based Radar-Camera Fusion
Xiaomeng Chu, Jiajun Deng, Guoliang You, Yifan Duan, Houqiang Li, Yanyong Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17081-17091
We propose Radar-Camera fusion transformer (RaCFormer) to boost the accuracy of 3D object detection by the following insight. The Radar-Camera fusion in outdoor 3D scene perception is capped by the image-to-BEV transformation-if the depth of pixels is not accurately estimated, the naive combination of BEV features actually integrates unaligned visual content. To avoid this problem, we propose a query-based framework that enables adaptive sampling of instance-relevant features from both the bird's-eye view (BEV) and the original image view. Furthermore, we enhance system performance by two key designs: optimizing query initialization and strengthening the representational capacity of BEV. For the former, we introduce an adaptive circular distribution in polar coordinates to refine the initialization of object queries, allowing for a distance-based adjustment of query density. For the latter, we initially incorporate a radar-guided depth head to refine the transformation from image view to BEV. Subsequently, we focus on leveraging the Doppler effect of radar and introduce an implicit dynamic catcher to ca

pture the temporal elements within the BEV. Extensive experiments on nuScenes an
d View-of-Delft (VoD) datasets validate the merits of our design. Remarkably, ou
r method achieves superior results of 64.9% mAP and 70.2% NDS on nuScenes. RaCFo
rmer also secures the state-of-the-art performance on the VoD dataset. Code is a
vailable at https://github.com/cxmomo/RaCFormer.
****

Hybrid Reciprocal Transformer with Triplet Feature Alignment for Scene Graph Gen
eration
Jiawei Fu, Tiantian Zhang, Kai Chen, Qi Dou; Proceedings of the Computer Vision
and Pattern Recognition Conference (CVPR), 2025, pp. 8953-8963
Scene graph generation is a pivotal task in computer vision, focusing on compreh
ensive identification of visual relation tuples embedded within images. The adva
ncement of methods involving triplets has sought to enhance task performance by
integrating triplets as contextual features for more precise predicate identific
ation from component level. However, challenges remain due to interference from
multi-role objects in overlapping tuples within complex environments, which impa
irs the model's ability to distinguish and align specific triplet features for r
easoning diverse semantics of multi-role objects. To address these issues, we in
troduce a novel framework that incorporates a triplet alignment model into a hyb
rid reciprocal transformer architecture, starting from using triplet mask featur
es to guide the learning of component-level relation graphs. To effectively dist
inguish multi-role objects characterized by overlapping visual relation tuples,
we introduce a triplet alignment loss, which provides multi-role objects with al
igned features from triplet and helps customize them. Additionally, we explore t
he inherent connectivity between hybrid aligned triplet and component features t
hrough a bidirectional refinement module, which enhances feature interaction and
 reciprocal reinforcement. Experimental results demonstrate that our model achie
ves state-of-the-art performance on the Visual Genome and Action Genome datasets
, underscoring its effectiveness and adaptability. Project page: hq-sg.github.io
.
****

Understanding Multi-Task Activities from Single-Task Videos
Yuhan Shen, Ehsan Elhamifar; Proceedings of the Computer Vision and Pattern Reco
gnition Conference (CVPR), 2025, pp. 19120-19131
(MT-TAS), a novel paradigm that addresses the challenges of interleaved actions
when performing multiple tasks simultaneously. Traditional action segmentation m
odels, trained on single-task videos, struggle to handle task switches and compl
ex scenes inherent in multi-task scenarios. To overcome these challenges, our MT
-TAS approach synthesizes multi-task video data from single-task sources using o
ur Multi-task Sequence Blending and Segment Boundary Learning modules. Additiona
lly, we propose to dynamically isolate foreground and background elements within
 video frames, addressing the intricacies of object layouts in multi-task scenar
ios and enabling a new two-stage temporal action segmentation framework with For
eground-Aware Action Refinement. Also, we introduce the Multi-task Egocentric Ki
tchen Activities (MEKA) dataset, containing 12 hours of egocentric multi-task vi
deos, to rigorously benchmark MT-TAS models. Extensive experiments demonstrate t
hat our framework effectively bridges the gap between single-task training and m
ulti-task testing, advancing temporal action segmentation with state-of-the-art
performance in complex environments.
****

Co-Speech Gesture Video Generation with Implicit Motion-Audio Entanglement
Xinjie Li, Ziyi Chen, Xinlu Yu, Iek-Heng Chu, Peng Chang, Jing Xiao; Proceedings
 of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 113
84-11394
Co-speech gestures are essential to non-verbal communication, enhancing both the
 naturalness and effectiveness of human interaction. Although recent methods hav
e made progress in generating co-speech gesture videos, many rely on strong visu
al controls, such as pose images or TPS keypoint movements, which often lead to
artifacts like blurry hands and distorted fingers. In response to these challeng
es, we present the Implicit Motion-Audio Entanglement (IMAE) method for co-speec

h gesture video generation. IMAE strengthens audio control by entangling implici t motion parameters, including pose and expression, with audio inputs. Our metho d utilizes a two-branch framework that combines an audio-to-motion generation br anch with a video diffusion branch, enabling realistic gesture generation withou t requiring additional inputs during inference. To improve training efficiency, we propose a two-stage slow-fast training strategy that balances memory constrai nts while facilitating the learning of meaningful gestures from long frame seque nces.Extensive experimental results demonstrate that our method achieves state-o f-the-art performance across multiple metrics.
****

TransPixeler: Advancing Text-to-Video Generation with Transparency
Luozhou Wang, Yijun Li, Zhifei Chen, Jui-Hsien Wang, Zhifei Zhang, He Zhang, Zhe Lin, Ying-Cong Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18229-18239
Text-to-video generative models have made significant strides, enabling diverse applications in entertainment, advertising, and education. However, generating R GBA video, which includes alpha channels for transparency, remains a challenge d ue to limited datasets and the difficulty of adapting existing models. Alpha cha nnels are crucial for visual effects (VFX), allowing transparent elements like s moke and reflections to blend seamlessly into scenes.We introduce TransPixeler, a method to extend pretrained video models for RGBA generation while retaining t he original RGB capabilities. TransPixar leverages a diffusion transformer (DiT) architecture, incorporating alpha-specific tokens and using LoRA-based fine-tun ing to jointly generate RGB and alpha channels with high consistency. By optimiz ing attention mechanisms, TransPixeler preserves the strengths of the original R GB model and achieves strong alignment between RGB and alpha channels despite li mited training data.Our approach effectively generates diverse and consistent RG BA videos, advancing the possibilities for VFX and interactive content creation.
****

Adaptive Keyframe Sampling for Long Video Understanding
Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, Qixiang Ye; Proceedin gs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2 9118-29128
Multimodal large language models (MLLMs) have enabled open-world visual understa nding by injecting visual input as extra tokens into large language models (LLMs) as contexts. However, when the visual input changes from a single image to a l ong video, the above paradigm encounters difficulty because the vast amount of v ideo tokens has significantly exceeded the maximal capacity of MLLMs. Therefore, existing video-based MLLMs are mostly established upon sampling a small portion of tokens from input data, which can cause key information to be lost and thus produce incorrect answers. This paper presents a simple yet effective algorithm named Adaptive Keyframe Sampling (AKS). It inserts a plug-and-play module known as keyframe selection, which aims to maximize the useful information with a fixe d number of video tokens. We formulate keyframe selection as an optimization inv olving (1) the relevance between the keyframes and the prompt, and (2) the cover age of the keyframes over the video, and present an adaptive algorithm to approx imate the best solution. Experiments on two long video understanding benchmarks validate that AKS improves video QA accuracy (beyond strong baselines) upon sele cting informative keyframes. Our study reveals the importance of information pre -filtering in video-based MLLMs. Our code are available at https://github.com/nc TimTang/AKS
****

What's in the Image? A Deep-Dive into the Vision of Vision Language Models
Omri Kaduri, Shai Bagon, Tali Dekel; Proceedings of the Computer Vision and Patt ern Recognition Conference (CVPR), 2025, pp. 14549-14558
Vision-Language Models (VLMs) have recently demonstrated remarkable capabilities in comprehending complex visual content. However, the mechanisms underlying how VLMs process visual information remain largely unexplored. In this paper, we c onduct a thorough empirical analysis, focusing on the attention modules across l ayers, by which we reveal several key insights about how these models process vi

sual data: (i) the internal representation of the query tokens (e.g., representations of "describe the image"), is utilized by the model to store global image information; we demonstrate that the model generates surprisingly descriptive responses solely from these tokens, without direct access to image tokens. (ii) Cross-modal information flow is predominantly influenced by the middle layers (approximately 25% of all layers), while early and late layers contribute only marginally. (iii) Fine-grained visual attributes and object details are directly extracted from image tokens in a spatially localized manner, i.e., the generated tokens associated with a specific object or attribute attend strongly to their corresponding regions in the image. We propose novel quantitative evaluation to validate our observations, leveraging real-world complex visual scenes. Finally, we demonstrate the potential of our findings in facilitating efficient visual processing in state-of-the-art VLMs.

****

Person De-reidentification: A Variation-guided Identity Shift Modeling

Yi-Xing Peng, Yu-Ming Tang, Kun-Yu Lin, Qize Yang, Jingke Meng, Xihan Wei, Wei-Shi Zheng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29331-29341

Person re-identification (ReID) is to associate images of individuals from different camera views against cross-view variations. Like other surveillance technologies, Re-ID faces serious privacy challenges, particularly the potential for unauthorized tracking. Although various tasks (e.g., face recognition) have developed machine unlearning techniques to address privacy concerns, such methods have not yet been explored within the Re-ID field. In this work, we pioneer the exploration of the person de-reidentification (De-ReID) problem and present its inherent challenges. In the context of ReID, De-ReID is to unlearn the knowledge about accurately matching specific persons so that these "unlearned persons" cannot be re-identified across cameras for privacy guarantee. The primary challenge is to achieve the unlearning without degrading the identity-discriminative feature embeddings to ensure the model's utility. To address this, we formulate a De-ReID framework that utilizes a labeled dataset of unlearned persons for unlearning and an unlabeled dataset of accessible persons for knowledge preservation. Instead of unlearning based on (pseudo) identity labels, we introduce a variation-guided identity shift mechanism that unlearns the specific persons by fitting the variations in their images while preserving ReID ability on other persons by overcoming the variations in images of accessible persons. As a result, the model shifts the unlearned persons to a feature space that is vulnerable to cross-view variations. Extensive experiments on benchmarks demonstrate the superiority of our method.

****

FreeSim: Toward Free-viewpoint Camera Simulation in Driving Scenes

Lue Fan, Hao Zhang, Qitai Wang, Hongsheng Li, Zhaoxiang Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12004-12014

We propose FreeSim, a camera simulation method for driving scenes via 3D Gaussian Splatting and diffusion-based image generation. FreeSim emphasizes high-quality rendering from viewpoints beyond the recorded ego trajectories. In such viewpoints, previous methods have unacceptable degradation because the training data of these viewpoints is unavailable. To address such data scarcity, we first propose a generative enhancement model with a matched data construction strategy. The resulting model can generate high-quality images in a viewpoint slightly deviated from the recorded trajectories, conditioned on the degraded rendering of this viewpoint. We then propose a progressive reconstruction strategy, which progressively adds generated images in unrecorded views into the reconstruction process, starting from slightly off-trajectory viewpoints and moving progressively farther away. With this progressive generation-reconstruction pipeline, FreeSim supports high-quality off-trajectory view synthesis under large deviations of more than 3 meters.

****

Gradient Inversion Attacks on Parameter-Efficient Fine-Tuning

Hasin Us Sami, Swapneel Sen, Amit K. Roy-Chowdhury, Srikanth V. Krishnamurthy, Basak Guler; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10224-10234

Federated learning (FL) allows multiple data-owners to collaboratively train machine learning models by exchanging local gradients, while keeping their private data on-device. To simultaneously enhance privacy and training efficiency, recently parameter-efficient fine-tuning (PEFT) of large-scale pretrained models has gained substantial attention in FL. While keeping a pretrained (backbone) model frozen, each user fine-tunes only a few lightweight modules to be used in conjunction, to fit specific downstream applications. Accordingly, only the gradients with respect to these lightweight modules are shared with the server. In this work, we investigate how the privacy of the fine-tuning data of the users can be compromised via a malicious design of the pretrained model and trainable adapter modules. We demonstrate gradient inversion attacks on a popular PEFT mechanism, the adapter, which allow an attacker to reconstruct local data samples of a target user, using only the accessible adapter gradients. Via extensive experiments, we demonstrate that a large batch of fine-tuning images can be retrieved with high fidelity. Our attack highlights the need for privacy-preserving mechanisms for PEFT, while opening up several future directions. Our code is available at https://github.com/info-ucr/PEFTLeak.

****

UPME: An Unsupervised Peer Review Framework for Multimodal Large Language Model Evaluation

Qihui Zhang, Munan Ning, Zheyuan Liu, Yue Huang, Shuo Yang, Yanbo Wang, Jiayi Ye, Xiao Chen, Yibing Song, Li Yuan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9165-9174

Multimodal Large Language Models (MLLMs) have emerged to tackle the challenges of Visual Question Answering (VQA), sparking a new research focus on conducting objective evaluations of these models. Existing evaluation mechanisms face limitations due to the significant human workload required to design Q&A pairs for visual images, which inherently restricts the scale and scope of evaluations. Although automated MLLM-as-judge approaches attempt to reduce the human workload through mutual model evaluations, they often introduce biases. To address these problems, we propose an unsupervised evaluation method -- an Unsupervised Peer review MLLM Evaluation framework. This framework utilizes only image data, allowing models to automatically generate questions and conduct peer review assessments of answers from other models, effectively alleviating the reliance on human workload. Additionally, we introduce the vision-language scoring system to mitigate the bias issues, which focuses on three aspects: (i) response correctness; (ii) the model capability of visual understanding and reasoning; (iii) relevance of text-image matching. Experimental results demonstrate that UPME achieves a Pearson correlation of 0.944 with human evaluations on the MMstar dataset and 0.814 on the ScienceQA dataset, indicating that our UPME framework closely aligns with human-designed QA benchmarks and inherent human preferences.

****

DiGIT: Multi-Dilated Gated Encoder and Central-Adjacent Region Integrated Decoder for Temporal Action Detection Transformer

Ho-Joong Kim, Yearang Lee, Jung-Ho Hong, Seong-Whan Lee; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24286-24296

In this paper, we examine a key limitation in query-based detectors for temporal action detection (TAD), which arises from their direct adaptation of originally designed architectures for object detection. Despite the effectiveness of the existing models, they struggle to fully address the unique challenges of TAD, such as the redundancy in multi-scale features and the limited ability to capture sufficient temporal context. To address these issues, we propose a multi-dilated gated encoder and central-adjacent region integrated decoder for temporal action detection transformer (DiGIT). Our approach replaces the existing encoder that consists of multi-scale deformable attention and feedforward network with our multi-dilated gated encoder. Our proposed encoder reduces the redundant information caused by multi-level features while maintaining the ability to capture fine-g

rained and long-range temporal information. Furthermore, we introduce a central-adjacent region integrated decoder that leverages a more comprehensive sampling strategy for deformable cross-attention to capture the essential information. Extensive experiments demonstrate that DiGIT achieves state-of-the-art performance on THUMOS14, ActivityNet v1.3, and HACS-Segment.
****

MBQ: Modality-Balanced Quantization for Large Vision-Language Models
Shiyao Li, Yingchun Hu, Xuefei Ning, Xihui Liu, Ke Hong, Xiaotao Jia, Xiuhong Li, Yaqi Yan, Pei Ran, Guohao Dai, Shengen Yan, Huazhong Yang, Yu Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4167-4177
Vision-Language Models (VLMs) have already enabled a variety of real-world applications. The large parameter size of VLMs brings large memory and computation overhead which poses significant challenges for deployment. Post-Training Quantization (PTQ) is an effective technique to reduce the memory and computation overhead. Existing PTQ methods mainly focus on the language modality in large language models (LLMs), without considering the differences across other modalities. In this paper, we discover that there is a significant difference in sensitivity between language and vision tokens in large VLMs. Therefore, treating tokens from different modalities equally, as in existing PTQ methods, may over-emphasize the insensitive modalities, leading to significant accuracy loss. To deal with the above issue, we propose a simple yet effective method, Modality-Balanced Quantization (MBQ), for large VLMs. Specifically, MBQ incorporates the different sensitivities across modalities during the calibration process to minimize the reconstruction loss for better quantization parameters. Extensive experiments show that MBQ can significantly improve task accuracy by up to 4.4% and 11.6% under W3A16 and W4A8 quantization for 7B to 70B VLMs, compared to SOTA baselines. Additionally, we implement a W3A16 GPU kernel that fuses the dequantization and GEMV operators, achieving a 1.4x speedup on LLaVA-onevision-7B on the RTX 4090. We will release the code.
****

Florence-VL: Enhancing Vision-Language Models with Generative Vision Encoder and Depth-Breadth Fusion
Jiuhai Chen, Jianwei Yang, Haiping Wu, Dianqi Li, Jianfeng Gao, Tianyi Zhou, Bin Xiao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24928-24938
We present Florence-VL, a new family of multimodal large language models (MLLMs) with enriched visual representations produced by Florence-2, a generative vision foundation model. Unlike the widely used CLIP-style vision transformer trained by contrastive learning, Florence-2 can capture different levels and aspects of visual features, which are more versatile to be adapted to diverse downstream tasks. We propose a novel feature-fusion architecture and an innovative training recipe that effectively integrates Florence-2's visual features into pretrained LLMs, such as Phi 3.5 and LLama 3. In particular, we propose "depth-breath fusion (DBFusion)" to fuse the visual features extracted from different depths and under multiple prompts. Our model training is composed of end-to-end pretraining of the whole model followed by finetuning of the projection layer and the LLM, on a carefully designed recipe of diverse open-source datasets that include high-quality image captions and instruction-tuning pairs. Our quantitative analysis and visualization of Florence-VL's visual features show its advantages over popular vision encoders on vision-language alignment, where the enriched depth and breath play important roles. Florence-VL achieves significant improvements over existing state-of-the-art MLLMs across various multi-modal and vision-centric benchmarks covering general VQA, perception, hallucination, OCR, Chart, knowledge-intensive understanding, etc. To facilitate future research, our models and the complete training recipe are open-sourced.
****

VideoDPO: Omni-Preference Alignment for Video Diffusion Generation
Runtao Liu, Haoyu Wu, Ziqiang Zheng, Chen Wei, Yingqing He, Renjie Pi, Qifeng Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)

, 2025, pp. 8009-8019
Recent progress in generative diffusion models has greatly advanced text-to-video generation. While text-to-video models trained on large-scale, diverse datasets can produce varied outputs, these generations often deviate from user preferences, highlighting the need for preference alignment on pre-trained models. Although Direct Preference Optimization (DPO) has demonstrated significant improvements in language and image generation, we pioneer its adaptation to video diffusion models and propose a VideoDPO pipeline by making several key adjustments. Unlike previous image alignment methods that focus solely on either (i) visual quality or (ii) semantic alignment between text and videos, we comprehensively consider both dimensions and construct a preference score accordingly, which we term the OmniScore. We design a pipeline to automatically collect preference pair data based on the proposed OmniScore and discover that re-weighting these pairs based on the score significantly impacts overall preference alignment. Our experiments demonstrate substantial improvements in both visual quality and semantic alignment, ensuring that no preference aspect is neglected. Code and data will be shared at https://videodpo.github.io/.
****

Seq2Time: Sequential Knowledge Transfer for Video LLM Temporal Grounding
Andong Deng, Zhongpai Gao, Anwesa Choudhuri, Benjamin Planche, Meng Zheng, Bin Wang, Terrence Chen, Chen Chen, Ziyan Wu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13766-13775
Temporal awareness is essential for video large language models (LLMs) to understand and reason about events within long videos, enabling applications like dense video captioning and temporal video grounding in a unified system. However, the scarcity of long videos with detailed captions and precise temporal annotations limits their temporal awareness. In this paper, we propose Seq2Time, a data-oriented training paradigm that leverages sequences of images and short video clips to enhance temporal awareness in long videos. By converting sequence positions into temporal annotations, we transform large-scale image and clip captioning datasets into sequences that mimic the temporal structure of long videos, enabling self-supervised training with abundant time-sensitive data. To enable sequence-to-time knowledge transfer, we introduce a novel time representation that unifies positional information across image sequences, clip sequences, and long videos. Experiments demonstrate the effectiveness of our method, achieving a 27.6% improvement in F1 score and 44.8% in CIDEr on the YouCook2 benchmark and a 14.7% increase in recall on the Charades-STA benchmark compared to the baseline.
****

GPVK-VL: Geometry-Preserving Virtual Keyframes for Visual Localization under Large Viewpoint Changes
Yunxuan Li, Lei Fan, Xiaoying Xing, Jianxiong Zhou, Ying Wu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16728-16738
Visual localization, the task of determining the position and orientation of a camera, typically involves three core components: offline construction of a keyframe database, efficient online keyframes retrieval, and robust local feature matching. However, significant challenges arise when there are large viewpoint disparities between the query view and the database, such as attempting localization in a corridor previously build from an opposing direction. Intuitively, this issue can be addressed by synthesizing a set of virtual keyframes that cover all viewpoints. However, existing methods for synthesizing novel views to assist localization often fail to ensure geometric accuracy under large viewpoint changes. In this paper, we introduce a confidence-aware geometric prior into 2D Gaussian splatting to ensure the geometric accuracy of the scene. Then we can render novel views through the mesh with clear structures and accurate geometry, even under significant viewpoint changes, enabling the synthesis of a comprehensive set of virtual keyframes. Incorporating this geometry-preserving virtual keyframe database into the localization pipeline significantly enhances the robustness of visual localization.
****

Realistic Test-Time Adaptation of Vision-Language Models

Maxime Zanella, Clément Fuchs, Christophe De Vleeschouwer, Ismail Ben Ayed; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25103-25112

The zero-shot capabilities of Vision-Language Models (VLMs) have been widely leveraged to improve predictive performance. However, previous works on transductive or test-time adaptation (TTA) often make strong assumptions about the data distribution, such as the presence of all classes. Our work challenges these favorable deployment scenarios and introduces a more realistic evaluation framework, including (i) a variable number of effective classes for adaptation within a single batch, and (ii) non-i.i.d. batches of test samples in online adaptation settings. We provide comprehensive evaluations, comparisons, and ablation studies that demonstrate how current transductive or TTA methods for VLMs systematically compromise the models' initial zero-shot robustness across various realistic scenarios, favoring performance gains under advantageous assumptions about the test sample distributions. Furthermore, we introduce StatA, a versatile method that can handle a wide range of deployment scenarios, including those with a variable number of effective classes at test time. Our approach incorporates a novel regularization term designed specifically for VLMs, which acts as a statistical anchor preserving the initial text-encoder knowledge, particularly in low-data regimes. Code available at https://github.com/MaxZanella/StatA.

****

SelfSplat: Pose-Free and 3D Prior-Free Generalizable 3D Gaussian Splatting

Gyeongjin Kang, Jisang Yoo, Jihyeon Park, Seungtae Nam, Hyeonsoo Im, Sangheon Shin, Sangpil Kim, Eunbyung Park; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22012-22022

We propose SelfSplat, a novel 3D Gaussian Splatting model designed to perform pose-free and 3D prior-free generalizable 3D reconstruction from unposed multi-view images. These settings are inherently ill-posed due to the lack of ground-truth data, learned geometric information, and the need to achieve accurate 3D reconstruction without finetuning, making it difficult for conventional methods to achieve high-quality results. Our model addresses these challenges by effectively integrating explicit 3D representations with self-supervised depth and pose estimation techniques, resulting in reciprocal improvements in both pose accuracy and 3D reconstruction quality. Furthermore, we incorporate a matching-aware pose estimation network and a depth refinement module to enhance geometry consistency across views, ensuring more accurate and stable 3D reconstructions. To present the performance of our method, we evaluated it on large-scale real-world datasets, including RealEstate10K, ACID, and DL3DV. SelfSplat achieves superior results over previous state-of-the-art methods in both appearance and geometry quality, also demonstrates strong cross-dataset generalization capabilities. Extensive ablation studies and analysis also validate the effectiveness of our proposed methods.

****

Enhancing Virtual Try-On with Synthetic Pairs and Error-Aware Noise Scheduling

Nannan Li, Kevin J. Shih, Bryan A. Plummer; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21238-21247

Given an isolated garment image in a canonical product view and a separate image of a person, the virtual try-on task aims to generate a new image of the person wearing the target garment.Prior virtual try-on works face two major challenges in achieving this goal: a) the paired (human, garment) training data has limited availability; b) generating textures on the human that perfectly match that of the prompted garment is difficult, often resulting in distorted text and faded textures. Our work explores ways to tackle these issues through both synthetic data as well as model refinement. We introduce a garment extraction model that generates (human, synthetic garment) pairs from a single image of a clothed individual. The synthetic pairs can then be used to augment the training of virtual try-on. We also propose an Error-Aware Refinement-based Schrodinger Bridge (EARSB) that surgically targets localized generation errors for correcting the output of a base virtual try-on model. To identify likely errors, we propose a weakly-supervised error classifier that localizes regions for refinement, subsequently au

gmenting the Schrodinger Bridge's noise schedule with its confidence heatmap. Experiments on VITON-HD and DressCode-Upper demonstrate that our synthetic data augmentation enhances the performance of prior work, while EARSB improves the overall image quality. In user studies, our model is preferred by the users in an average of 59% of cases.

****

Exploring Simple Open-Vocabulary Semantic Segmentation

Zihang Lai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30221-30230

Open-vocabulary semantic segmentation models aim to accurately assign a semantic label to each pixel in an image from a set of arbitrary open-vocabulary texts. In order to learn such pixel-level alignment, current approaches typically rely on a combination of (i) image-level VL model (e.g. CLIP), (ii) ground truth masks, (iii) custom grouping encoders, and (iv) the Segment Anything Model (SAM). In this paper, we introduce S-Seg, a simple model that can achieve surprisingly strong performance without depending on any of the above elements. S-Seg leverages pseudo-mask and language to train a MaskFormer, and can be easily trained from publicly available image-text datasets. Contrary to prior works, our model directly trains for pixel-level features and language alignment. Once trained, S-Seg generalizes well to multiple testing datasets without requiring fine-tuning. In addition, S-Seg has the extra benefits of scalability with data and consistently improving when augmented with self-training. We believe that our simple yet effective approach will serve as a solid baseline for future research. Our code and demo will be made publicly available soon.

****

MP-GUI: Modality Perception with MLLMs for GUI Understanding

Ziwei Wang, Weizhi Chen, Leyang Yang, Sheng Zhou, Shengchu Zhao, Hanbei Zhan, Jiongchao Jin, Liangcheng Li, Zirui Shao, Jiajun Bu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29711-29721

Graphical user interface (GUI) has become integral to modern society, making it crucial to be understood for human-centric systems. However, unlike natural images or documents, GUIs comprise artificially designed graphical elements arranged to convey specific semantic meanings. Current multi-modal large language models (MLLMs) already proficient in processing graphical and textual components suffer from hurdles in GUI understanding due to the lack of explicit spatial structure modeling. Moreover, obtaining high-quality spatial structure data is challenging due to privacy issues and noisy environments. To address these challenges, we present MP-GUI, a specially designed MLLM for GUI understanding. MP-GUI features three precisely specialized perceivers to extract graphical, textual, and spatial modalities from the screen as GUI-tailored visual clues, with spatial structure refinement strategy and adaptively combined via a fusion gate to meet the specific preferences of different GUI understanding tasks. To cope with the scarcity of training data, we also introduce a pipeline for automatically data collecting. Extensive experiments demonstrate that MP-GUI achieves impressive results on various GUI understanding tasks with limited data. Our codes and datasets are publicly available at https://github.com/BigTaige/MP-GUI.

****

Associative Transformer

Yuwei Sun, Hideya Ochiai, Zhirong Wu, Stephen Lin, Ryota Kanai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4518-4527

Emerging from the pairwise attention in conventional Transformers, there is a growing interest in sparse attention mechanisms that align more closely with localized, contextual learning in the biological brain. Existing studies such as the Coordination method employ iterative cross-attention mechanisms with a bottleneck to enable the sparse association of inputs. However, these methods are parameter inefficient and fail in more complex relational reasoning tasks. To this end, we propose Associative Transformer (AiT) to enhance the association among sparsely attended input tokens, improving parameter efficiency and performance in various vision tasks such as classification and relational reasoning. AiT leverages

a learnable explicit memory comprising specialized priors that guide bottleneck attentions to facilitate the extraction of diverse localized tokens. Moreover, AiT employs an associative memory-based token reconstruction using a Hopfield energy function. The extensive empirical experiments demonstrate that AiT requires significantly fewer parameters and attention layers outperforming a broad range of sparse Transformer models. Additionally, AiT outperforms the SOTA sparse Transformer models including the Coordination method on the Sort-of-CLEVR dataset.
****

Improving Adversarial Transferability on Vision Transformers via Forward Propagation Refinement

Yuchen Ren, Zhengyu Zhao, Chenhao Lin, Bo Yang, Lu Zhou, Zhe Liu, Chao Shen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25071-25080

Vision Transformers (ViTs) have been widely applied in various computer vision and vision-language tasks. To gain insights into their robustness in practical scenarios, transferable adversarial examples on ViTs have been extensively studied. A typical approach to improving adversarial transferability is by refining the surrogate model. However, existing work on ViTs has restricted their surrogate refinement to backward propagation. In this work, we instead focus on Forward Propagation Refinement (FPR) and specifically refine two key modules of ViTs: attention maps and token embeddings. For attention maps, we propose Attention Map Diversification (AMD), which diversifies certain attention maps and also implicitly imposes beneficial gradient vanishing during backward propagation. For token embeddings, we propose Momentum Token Embedding (MTE), which accumulates historical token embeddings to stabilize the forward updates in both the Attention and MLP blocks. We conduct extensive experiments with adversarial examples transferred from ViTs to various CNNs and ViTs, demonstrating that our FPR outperforms the current best (backward) surrogate refinement by up to 7.0% on average. We also validate its superiority against popular defenses and its compatibility with other transfer methods. Codes and appendix are available at https://github.com/RYC-98/FPR.
****

Seeing What Matters: Empowering CLIP with Patch Generation-to-Selection

Gensheng Pei, Tao Chen, Yujia Wang, Xinhao Cai, Xiangbo Shu, Tianfei Zhou, Yazhou Yao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24862-24872

The CLIP model has demonstrated significant advancements in aligning visual and language modalities through large-scale pre-training on image-text pairs, enabling strong zero-shot classification and retrieval capabilities on various domains. However, CLIP's training remains computationally intensive, with high demands on both data processing and memory. To address these challenges, recent masking strategies have emerged, focusing on the selective removal of image patches to improve training efficiency. Although effective, these methods often compromise key semantic information, resulting in suboptimal alignment between visual features and text descriptions.In this work, we present a concise yet effective approach called Patch Generation-to-Selection (CLIP-PGS) to enhance CLIP's training efficiency while preserving critical semantic content. Our method introduces a gradual masking process in which a small set of candidate patches is first pre-selected as potential mask regions. Then, we apply Sobel edge detection across the entire image to generate an edge mask that prioritizes the retention of the primary object areas. Finally, similarity scores between the candidate mask patches and their neighboring patches are computed, with optimal transport normalization refining the selection process to ensure a balanced similarity matrix.Our approach, CLIP-PGS, sets new state-of-the-art results in zero-shot classification and retrieval tasks, achieving superior performance in robustness evaluation and language compositionality benchmarks.
****

ChatGarment: Garment Estimation, Generation and Editing via Large Language Models

Siyuan Bian, Chenghao Xu, Yuliang Xiu, Artur Grigorev, Zhen Liu, Cewu Lu, Michae

l J. Black, Yao Feng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2924-2934

We introduce ChatGarment, a novel approach that leverages large vision-language models (VLMs) to automate the estimation, generation, and editing of 3D garment sewing patterns from images or text descriptions. Unlike previous methods that often lack robustness and interactive editing capabilities, ChatGarment finetunes a VLM to produce GarmentCode, a JSON-based, language-friendly format for 2D sewing patterns, enabling both estimating and editing from images and text instructions. To optimize performance, we refine GarmentCode by expanding its support for more diverse garment types and simplifying its structure, making it more efficient for VLM finetuning. Additionally, we develop an automated data construction pipeline to generate a large-scale dataset of image-to-sewing-pattern and text-to-sewing-pattern pairs, empowering ChatGarment with strong generalization across various garment types. Extensive evaluations demonstrate ChatGarment's ability to accurately reconstruct, generate, and edit garments from multimodal inputs, highlighting its potential to revolutionize workflows in fashion and gaming applications.

****

Enhancing Online Continual Learning with Plug-and-Play State Space Model and Class-Conditional Mixture of Discretization

Sihao Liu, Yibo Yang, Xiaojie Li, David A. Clifton, Bernard Ghanem; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20502-20511

Online continual learning (OCL) seeks to learn new tasks from data streams that appear only once, while retaining knowledge of previously learned tasks. Most existing methods rely on replay, focusing on enhancing memory retention through regularization or distillation. However, they often overlook the adaptability of the model, limiting the ability to learn generalizable and discriminative features incrementally from online training data.To address this, we introduce a plug-and-play module, S6MOD, which can be integrated into most existing methods and directly improve adaptability. Specifically, S6MOD introduces an extra branch after the backbone, where a mixture of discretization selectively adjusts parameters in a selective state space model, enriching selective scan patterns such that the model can adaptively select the most sensitive discretization method for current dynamics.We further design a class-conditional routing algorithm for dynamic, uncertainty-based adjustment and implement a contrastive discretization loss to optimize it. Extensive experiments combining our module with various models demonstrate that S6MOD significantly enhances model adaptability, leading to substantial performance gains and achieving the state-of-the-art results. The code is available at https://github.com/MyToumaKazusa/S6MOD.

****

RDD: Robust Feature Detector and Descriptor using Deformable Transformer

Gonglin Chen, Tianwen Fu, Haiwei Chen, Wenbin Teng, Hanyuan Xiao, Yajie Zhao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6394-6403

As a core step in structure-from-motion and SLAM, robust feature detection and description under challenging scenarios such as significant viewpoint changes remain unresolved despite their ubiquity. While recent works have identified the importance of local features in modeling geometric transformations, these methods fail to learn the visual cues present in long-range relationships. We present Robust Deformable Detector (RDD), a novel and robust keypoint detector/descriptor leveraging the deformable transformer, which captures global context and geometric invariance through deformable self-attention mechanisms. Specifically, we observed that deformable attention focuses on key locations, effectively reducing the search space complexity and modeling the geometric invariance. Furthermore, we collected an Air-to-Ground dataset for training in addition to the standard MegaDepth dataset. Our proposed method outperforms all state-of-the-art keypoint detection/description methods in sparse matching tasks and is also capable of semi-dense matching. To ensure comprehensive evaluation, we introduce two challenging benchmarks: one emphasizing large viewpoint and scale variations, and the oth

er being an Air-to-Ground benchmark -- an evaluation setting that has recently g aining popularity for 3D reconstruction across different altitudes.
****

Building Vision Models upon Heat Conduction
Zhaozhi Wang, Yue Liu, Yunjie Tian, Yunfan Liu, Yaowei Wang, Qixiang Ye; Proceed ings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9707-9717
Visual representation models leveraging attention mechanisms are challenged by s ignificant computational overhead, particularly when pursuing large receptive fi elds. In this study, we aim to mitigate this challenge by introducing the Heat C onduction Operator (HCO) built upon the physical heat conduction principle. HCO conceptualizes image patches as heat sources and models their correlations throu gh adaptive thermal energy diffusion, enabling robust visual representations. HC O enjoys a computational complexity of O(N^1.5), as it can be implemented using discrete cosine transformation (DCT) operations. HCO is plug-and-play, combining with deep learning backbones produces visual representation models (termed vHea t) with global receptive fields. Experiments across vision tasks demonstrate tha t, beyond the stronger performance, vHeat achieves up to a 3x throughput, 80% le ss GPU memory allocation, and 35% fewer computational FLOPs compared to the Swin -Transformer.
****

GRAPHGPT-O: Synergistic Multimodal Comprehension and Generation on Graphs
Yi Fang, Bowen Jin, Jiacheng Shen, Sirui Ding, Qiaoyu Tan, Jiawei Han; Proceedin gs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1 9467-19476
The rapid development of Multimodal Large Language Models (MLLMs) has enabled th e integration of multiple modalities, including texts and images, within the lar ge language model (LLM) framework.However, texts and images are usually intercon nected, forming a multimodal attributed graph (MMAG).It is underexplored how MLL Ms can incorporate the relational information (i.e., graph structure) and semant ic information (i.e., texts and images) on such graphs for multimodal comprehens ion and generation.In this paper, we propose GraphGPT-o, which supports omni-mul timodal understanding and creation on MMAGs.We first comprehensively study linea rization variants to transform semantic and structural information as input for MLLMs.Then, we propose a hierarchical aligner that enables deep graph encoding, bridging the gap between MMAGs and MLLMs.Finally, we explore the inference choic es, adapting MLLM to interleaved text and image generation in graph scenarios. E xtensive experiments on three datasets from different domains demonstrate the ef fectiveness of our proposed method.
****

Model Poisoning Attacks to Federated Learning via Multi-Round Consistency
Yueqi Xie, Minghong Fang, Neil Zhenqiang Gong; Proceedings of the Computer Visio n and Pattern Recognition Conference (CVPR), 2025, pp. 15454-15463
Model poisoning attacks are critical security threats to Federated Learning (FL) . Existing model poisoning attacks suffer from two key limitations: 1) they achi eve suboptimal effectiveness when defenses are deployed, and/or 2) they require knowledge of the model updates or local training data on genuine clients. In thi s work, we make a key observation that their suboptimal effectiveness arises fro m only leveraging model-update consistency among malicious clients within indivi dual training rounds, making the attack effect self-cancel across training round s. In light of this observation, we propose PoisonedFL, which enforces multi-rou nd consistency among the malicious clients' model updates while not requiring an y knowledge about the genuine clients. Our empirical evaluation on five benchmar k datasets shows that PoisonedFL breaks eight state-of-the-art defenses and outp erforms seven existing model poisoning attacks. Our study shows that FL systems are considerably less robust than previously thought, underlining the urgency fo r the development of new defense mechanisms. Our source code is available at htt ps://github.com/xyq7/PoisonedFL/.
****

TaoAvatar: Real-Time Lifelike Full-Body Talking Avatars for Augmented Reality vi

a 3D Gaussian Splatting
Jianchuan Chen, Jingchuan Hu, Gaige Wang, Zhonghua Jiang, Tiansong Zhou, Zhiwen Chen, Chengfei Lv; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10723-10734
Realistic 3D full-body talking avatars hold great potential in AR, with applications ranging from e-commerce live streaming to holographic communication. Despite advances in 3D Gaussian Splatting (3DGS) for lifelike avatar creation, existing methods struggle with fine-grained control of facial expressions and body movements in full-body talking tasks. Additionally, they often lack sufficient details and cannot run in real-time on mobile devices. We present TaoAvatar, a high-fidelity, lightweight, 3DGS-based full-body talking avatar driven by various signals. Our approach starts by creating a personalized clothed human parametric template that binds Gaussians to represent appearances. We then pre-train a StyleUnet-based network to handle complex pose-dependent non-rigid deformation, which can capture high-frequency appearance details but is too resource-intensive for mobile devices. To overcome this, we "bake" the non-rigid deformations into a lightweight MLP-based network using a distillation technique and develop blend shapes to compensate for details. Extensive experiments show that TaoAvatar achieves state-of-the-art rendering quality while running in real-time across various devices, maintaining 90 FPS on high-definition stereo devices such as the Apple Vision Pro.
****
Erasing Undesirable Influence in Diffusion Models
Jing Wu, Trung Le, Munawar Hayat, Mehrtash Harandi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28263-28273
Diffusion models are highly effective at generating high-quality images but pose risks, such as the unintentional generation of NSFW (not safe for work) content.Although various techniques have been proposed to mitigate unwanted influences in diffusion models while preserving overall performance, achieving a balance between these goals remains challenging.In this work, we introduce EraseDiff, an algorithm designed to preserve the utility of the diffusion model on retained data while removing the unwanted information associated with the data to be forgotten.Our approach formulates this task as a constrained optimization problem using the value function, resulting in a natural first-order algorithm for solving the optimization problem.By altering the generative process to deviate away from the ground-truth denoising trajectory, we update parameters for preservation while controlling constraint reduction to ensure effective erasure, striking an optimal trade-off.Extensive experiments and thorough comparisons with state-of-the-art algorithms demonstrate that EraseDiff effectively preserves the model's utility, efficacy, and efficiency.
****
LT3SD: Latent Trees for 3D Scene Diffusion
Quan Meng, Lei Li, Matthias Nießner, Angela Dai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 650-660
We present LT3SD, a novel latent diffusion model for large-scale 3D scene generation. Recent advances in diffusion models have shown impressive results in 3D object generation, but are limited in spatial extent and quality when extended to 3D scenes. To generate complex and diverse 3D scene structures, we introduce a latent tree representation to effectively encode both lower-frequency geometry and higher-frequency detail in a coarse-to-fine hierarchy. We can then learn a generative diffusion process in this latent 3D scene space, modeling the latent components of a scene at each resolution level. To synthesize large-scale scenes with varying sizes, we train our diffusion model on scene patches and synthesize arbitrary-sized output 3D scenes through shared diffusion generation across multiple scene patches. Through extensive experiments, we demonstrate the efficacy and benefits of LT3SD for large-scale, high-quality unconditional 3D scene generation and for probabilistic completion for partial scene observations.
****
Stacking Brick by Brick: Aligned Feature Isolation for Incremental Face Forgery Detection

Jikang Cheng, Zhiyuan Yan, Ying Zhang, Li Hao, Jiaxin Ai, Qin Zou, Chen Li, Zhon gyuan Wang; Proceedings of the Computer Vision and Pattern Recognition Conferenc e (CVPR), 2025, pp. 13927-13936

The rapid advancement of face forgery techniques has introduced a growing variet y of forgeries.Incremental Face Forgery Detection (IFFD), involvinggradually add ing new forgery data to fine-tune the previously trained model, has been introdu ced as a promising strategy to deal with evolving forgery methods.However, a nai vely trained IFFD model is prone to catastrophic forgetting when new forgeries a re integrated, as treating all forgeries as a single "Fake" class in the Real/Fa ke classification can cause different forgery types overriding one another, ther eby resulting in the forgetting of unique characteristics from earlier tasks and limiting the model's effectiveness in learning forgery specificity and generali ty.In this paper, we propose to stack the latent feature distributions of previo us and new tasks brick by brick, i.e., achieving aligned feature isolation. In this manner, we aim to preserve learned forgery information and accumulate new k nowledge by minimizing distribution overriding, thereby mitigating catastrophic forgetting.To achieve this, we first introduce Sparse Uniform Replay (SUR) to ob tain the representative subsets that could be treated as the uniformly sparse ve rsions of the previous global distributions.We then propose a Latent-space Incre mental Detector (LID) that leverages SUR data to isolate and align distributions . For evaluation, we construct a more advanced and comprehensive benchmark tailo red for IFFD. The leading experimental results validate the superiority of our m ethod.

****

CO-SPY: Combining Semantic and Pixel Features to Detect Synthetic Images by AI

Siyuan Cheng, Lingjuan Lyu, Zhenting Wang, Xiangyu Zhang, Vikash Sehwag; Proceed ings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13455-13465

With the rapid advancement of generative AI, it is now possible to synthesize hi gh-quality images in a few seconds. Despite the power of these technologies, the y raise significant concerns regarding misuse. Current efforts to distinguish be tween real and AI-generated images may lack generalization, being effective for only certain types of generative models and susceptible to post-processing techn iques like JPEG compression. To overcome these limitations, we propose a novel f ramework, CO-SPY, that first enhances existing semantic features (e.g., the numb er of fingers in a hand) and artifact features (e.g., pixel value differences), and then adaptively integrates them to achieve more general and robust synthetic image detection. Additionally, we create CO-SPYBench, a comprehensive dataset c omprising 5 real image datasets and 22 state-of-the-art generative models, inclu ding the latest models like FLUX. We also collect 50k synthetic images in the wi ld from the Internet to enable evaluation in a more practical setting. Our exten sive evaluations demonstrate that our detector outperforms existing methods unde r identical training conditions, achieving an average accuracy improvement of ap proximately 11% to 34%.

****

Closest Neighbors are Harmful for Lightweight Masked Auto-encoders

Jian Meng, Ahmed Hasssan, Li Yang, Deliang Fan, Jinwoo Shin, Jae-sun Seo; Procee dings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp . 25230-25239

Learning the visual representation via masked auto-encoder (MAE) training has be en proven to be a powerful technique. Transferring the pre-trained vision transf ormer (ViT) to downstream tasks leads to superior performance compared to conven tional task-by-task supervised learning. Recent research works on MAE focus on large-sized vision transformers (>50 million parameters) with outstanding perfor mance. However, improving the generality of the under-parametrized lightweight m odel has been widely ignored. In practice, downstream applications are commonly intended for resource-constrained platforms, where large-scale ViT cannot easil y meet the resource budget. Current lightweight MAE training heavily relies on k nowledge distillation with a pre-trained teacher, whereas the root cause behind the poor performance remains under-explored. Motivated by that, this paper firs

t introduces the concept of "closest neighbor patch" to characterize the local s emantics among the input tokens. Our discovery shows that the lightweight model failed to distinguish different local information, leading to aliased understand ing and poor accuracy. Motivated by this finding, we propose NoR-MAE, a novel MA E training algorithm for lightweight vision transformers. NoR-MAE elegantly repe ls the semantic aliasing between patches and their closest neighboring patch (se mantic centroid) with negligible training cost overhead. With the ViT-Tiny model , NoR-MAE achieves up to 7.22%/3.64% accuracy improvements on ImageNet-100/Image Net-1K datasets, as well as up to 5.13% accuracy improvements in tested downstre am tasks. https://github.com/SeoLabCornell/NoR-MAE
****

CraftsMan3D: High-fidelity Mesh Generation with 3D Native Diffusion and Interact ive Geometry Refiner
Weiyu Li, Jiarui Liu, Hongyu Yan, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, Xiaoxiao Long; Proceedings of the Computer Vision and Pattern Recognition Confer ence (CVPR), 2025, pp. 5307-5317
We present a novel generative 3D modeling system, coined CraftsMan, which can ge nerate high-fidelity 3D geometries with highly varied shapes, regular mesh topol ogies, and detailed surfaces, and, notably, allows for refining the geometry in an interactive manner. Despite the significant advancements in 3D generation, ex isting methods still struggle with lengthy optimization processes, self-occlusio n, irregular mesh topologies, and difficulties in accommodating user edits, cons equently impeding their widespread adoption and implementation in 3D modeling so ftwares. Our work is inspired by the craftsman, who usually roughs out the holis tic figure of the work first and elaborates the surface details subsequently. Sp ecifically, we first introduce a robust data preprocessing pipeline that utilize s visibility check and winding mumber to maximize the use of existing 3D data. L everaging this data, we employ a 3D-native DiT model that directly models the di stribution of 3D data in latent space, generating coarse geometries with regular mesh topology in seconds. Subsequently, a normal-based geometry refiner enhance s local surface details, which can be applied automatically or interactively wit h user input. Extensive experiments demonstrate that our method achieves high ef ficacy in producing superior quality 3D assets compared to existing methods.
****

Decouple-Then-Merge: Finetune Diffusion Models as Multi-Task Learning
Qianli Ma, Xuefei Ning, Dongrui Liu, Li Niu, Linfeng Zhang; Proceedings of the C omputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23281-23291
Diffusion models are trained by learning a sequence of models that reverse each step of noise corruption. Typically, the model parameters are fully shared acros s multiple timesteps to enhance training efficiency. However, since the denoisin g tasks differ at each timestep, the gradients computed at different timesteps m ay conflict, potentially degrading the overall performance of image generation. To solve this issue, this work proposes a Decouple-then-Merge (DeMe) framework, which begins with a pretrained model and finetunes separate models tailored to s pecific timesteps. We introduce several improved techniques during the finetunin g stage to promote effective knowledge sharing while minimizing training interfe rence across timesteps. Finally, after finetuning, these separate models can be merged into a single model in the parameter space, ensuring efficient and practi cal inference. Experimental results show significant generation quality improvem ents upon 6 benchmarks including Stable Diffusion on COCO30K, ImageNet1K, PartiP rompts, and DDPM on LSUN Church, LSUN Bedroom, and CIFAR10. Code is included in the supplementary material and will be released on Github.
****

GIF: Generative Inspiration for Face Recognition at Scale
Saeed Ebrahimi, Sahar Rahimi, Ali Dabouei, Srinjoy Das, Jeremy M. Dawson, Nasser M. Nasrabadi; Proceedings of the Computer Vision and Pattern Recognition Confer ence (CVPR), 2025, pp. 3528-3539
Aiming to reduce the computational cost of Softmax in massive label space of Fac e Recognition (FR) benchmarks, recent studies estimate the output using a subset of identities. Although promising, the association between the computation cost

and the number of identities in the dataset remains linear only with a reduced ratio. A shared characteristic among available FR methods is the employment of atomic scalar labels during training. Consequently, the input to label matching is through a dot product between the feature vector of the input and the Softmax centroids. Inspired by generative modeling, we present a simple yet effective method that substitutes scalar labels with structured identity code, i.e., a sequence of integers. Specifically, we propose a tokenization scheme that transforms atomic scalar labels into structured identity codes. Then, we train an FR backbone to predict the code for each input instead of its scalar label. As a result, the associated computational cost becomes logarithmic \wrt number of identities. We demonstrate the benefits of the proposed method by conducting experiments. In particular, our method outperforms its competitors by 1.52%, and 0.6% at TAR@FAR=1e-4 on IJB-B and IJB-C, respectively, while transforming the association between computational cost and the number of identities from linear to logarithmic. \href https://github.com/msed-Ebrahimi/GIF Code

****

HELVIPAD: A Real-World Dataset for Omnidirectional Stereo Depth Estimation

Mehdi Zayene, Jannik Endres, Albias Havolli, Charles Corbière, Salim Cherkaoui, Alexandre Kontouli, Alexandre Alahi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26975-26984

Despite progress in stereo depth estimation, omnidirectional imaging remains underexplored, mainly due to the lack of appropriate data. We introduce Helvipad, a real-world dataset for omnidirectional stereo depth estimation, featuring 40K video frames from video sequences across diverse environments, including crowded indoor and outdoor scenes with various lighting conditions. Collected using two 360deg cameras in a top-bottom setup and a LiDAR sensor, the dataset includes accurate depth and disparity labels by projecting 3D point clouds onto equirectangular images. Additionally, we provide an augmented training set with an increased label density by using depth completion. We benchmark leading stereo depth estimation models for both standard and omnidirectional images. The results show that while recent stereo methods perform decently, a challenge persists in accurately estimating depth in omnidirectional imaging. To address this, we introduce necessary adaptations to stereo models, leading to improved performance.

****

GENMANIP: LLM-driven Simulation for Generalizable Instruction-Following Manipulation

Ning Gao, Yilun Chen, Shuai Yang, Xinyi Chen, Yang Tian, Hao Li, Haifeng Huang, Hanqing Wang, Tai Wang, Jiangmiao Pang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12187-12198

Robotic manipulation in real-world settings remains challenging, especially regarding robust generalization. Existing simulation platforms lack sufficient support for exploring how policies adapt to varied instructions and scenarios. Thus, they lag behind the growing interest in instruction-following foundation models like LLMs, whose adaptability is crucial yet remains underexplored in fair comparisons. To bridge this gap, we introduce GenManip, a realistic tabletop simulation platform tailored for policy generalization studies. It features an automatic pipeline via LLM-driven task-oriented scene graph to synthesize large-scale, diverse tasks using 10K annotated 3D object assets. To systematically assess generalization, we present GenManip-Bench, a benchmark of 200 scenarios refined via human-in-the-loop corrections. We evaluate two policy types: (1) modular manipulation systems integrating foundation models for perception, reasoning, and planning, and (2) end-to-end policies trained through scalable data collection. Results show that while data scaling benefits end-to-end methods, modular systems enhanced with foundation models generalize more effectively across diverse scenarios. We anticipate this platform to facilitate critical insights for advancing policy generalization in realistic conditions. All code will be made publicly available.

****

SKDream: Controllable Multi-view and 3D Generation with Arbitrary Skeletons

Yuanyou Xu, Zongxin Yang, Yi Yang; Proceedings of the Computer Vision and Patter

n Recognition Conference (CVPR), 2025, pp. 314-325
Controllable generation has achieved substantial progress in both 2D and 3D doma ins, yet current conditional generation methods still face limitations in descri bing detailed shape structures. Skeletons can effectively represent and describe object anatomy and pose. Unfortunately, past studies are often limited to human skeletons. In this work, we generalize skeletal conditioned generation to arbit rary structures. First, we design a reliable mesh skeletonization pipeline to ge nerate a large-scale mesh-skeleton paired dataset.Based on the dataset, a multi- view and 3D generation pipeline is built. We propose to represent 3D skeletons b y Coordinate Color Encoding as 2D conditional images. A Skeletal Correlation Mod ule is designed to extract global skeletal features for condition injection. Aft er multi-view images are generated, 3D assets can be obtained by incorporating a large reconstruction model, followed by a UV texture refinement stage. As a res ult, our method achieves instant generation of multi-view and 3D contents that a re aligned with given skeletons. The proposed techniques largely improve the obj ect-skeleton alignment and generation quality.
****

Towards Enhanced Image Inpainting: Mitigating Unwanted Object Insertion and Pres erving Color Consistency
Yikai Wang, Chenjie Cao, Junqiu Yu, Ke Fan, Xiangyang Xue, Yanwei Fu; Proceeding s of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23 237-23248
Recent advances in image inpainting increasingly use generative models to handle large irregular masks. However, these models can create unrealistic inpainted i mages due to two main issues: (1) Unwanted object insertion: Even with unmasked areas as context, generative models may still generate arbitrary objects in the masked region that don't align with the rest of the image. (2) Color inconsisten cy: Inpainted regions often have color shifts that causes a smeared appearance, reducing image quality. Retraining the generative model could help solve these i ssues, but it's costly since state-of-the-art latent-based diffusion and rectifi ed flow models require a three-stage training process: training a VAE, training a generative U-Net or transformer, and fine-tuning for inpainting. Instead, this paper proposes a post-processing approach, dubbed as ASUKA (Aligned Stable inpa inting with UnKnown Areas prior), to improve inpainting models. To address unwan ted object insertion, we leverage a Masked Auto-Encoder (MAE) for reconstruction -based priors. This mitigates object hallucination while maintaining the model's generation capabilities. To address color inconsistency, we propose a specializ ed VAE decoder that treats latent-to-image decoding as a local harmonization tas k, significantly reducing color shifts for color-consistent inpainting. We valid ate ASUKA on SD 1.5 and FLUX inpainting variants with Places2 and MISATO, our pr oposed diverse collection of datasets. Results show that ASUKA mitigates object hallucination and improves color consistency over standard diffusion and rectifi ed flow models and other inpainting methods.
****

Optimus-2: Multimodal Minecraft Agent with Goal-Observation-Action Conditioned P olicy
Zaijing Li, Yuquan Xie, Rui Shao, Gongwei Chen, Dongmei Jiang, Liqiang Nie; Proc eedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9039-9049
Building an agent that can mimic human behavior patterns to accomplish various o pen-world tasks is a long-term goal. To enable agents to effectively learn beha vioral patterns across diverse tasks, a key challenge lies in modeling the intri cate relationships among observations, actions, and language. To this end, we pr opose Optimus-2, a novel Minecraft agent that incorporates a Multimodal Large La nguage Model (MLLM) for high-level planning, alongside a Goal-Observation-Action Conditioned Policy (GOAP) for low-level control. GOAP contains (1) an Action-gu ided Behavior Encoder that models causal relationships between observations and actions at each timestep, then dynamically interacts with the historical observa tion-action sequence, consolidating it into fixed-length behavior tokens, and (2) an MLLM that aligns behavior tokens with open-ended language instructions to p

redict actions auto-regressively. Moreover, we introduce a high-quality Minecraft Goal-Observation-Action (MGOA) dataset, which contains 25,000 videos across 8 atomic tasks, providing about 30M goal-observation-action pairs. The automated construction method, along with the MGOA dataset, can contribute to the community's efforts in training Minecraft agents. Extensive experimental results demonstrate that Optimus-2 exhibits superior performance across atomic tasks, long-horizon tasks, and open-ended instruction tasks in Minecraft.
****

Classic Video Denoising in a Machine Learning World: Robust, Fast, and Controllable

Xin Jin, Simon Niklaus, Zhoutong Zhang, Zhihao Xia, Chunle Guo, Yuting Yang, Jiawen Chen, Chongyi Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2084-2093

Denoising is a crucial step in many video processing pipelines such as in interactive editing, where high quality, speed, and user control are essential. While recent approaches achieve significant improvements in denoising quality by leveraging deep learning, they are prone to unexpected failures due to discrepancies between training data distributions and the wide variety of noise patterns found in real-world videos. These methods also tend to be slow and lack user control. In contrast, traditional denoising methods perform reliably on in-the-wild videos and run relatively quickly on modern hardware. However, they require manually tuning parameters for each input video, which is not only tedious but also requires skill. We bridge the gap between these two paradigms by proposing a differentiable denoising pipeline based on traditional methods. A neural network is then trained to predict the optimal denoising parameters for each specific input, resulting in a robust and efficient approach that also supports user control.
****

Practical Solutions to the Relative Pose of Three Calibrated Cameras

Charalambos Tzamos, Viktor Kocur, Yaqing Ding, Daniel Barath, Zuzana Berger Haladova, Torsten Sattler, Zuzana Kukelova; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21913-21923

We study the challenging problem of estimating the relative pose of three calibrated cameras from four point correspondences. We propose novel efficient solutions to this problem that are based on the simple idea of using four correspondences to estimate an approximate geometry of the first two views. We model this geometry either as an affine or a fully perspective geometry estimated using one additional approximate correspondence. We generate such an approximate correspondence using a very simple and efficient strategy, where the new point is the mean point of three corresponding input points. The new solvers are efficient and easy to implement, since they are based on existing efficient minimal solvers, i.e., the 4-point affine fundamental matrix, the well-known 5-point relative pose solver, and the \texttt P3P solver. Extensive experiments on real data show that the proposed solvers, when properly coupled with local optimization, achieve state-of-the-art results, with the novel solver based on approximate mean-point correspondences being more robust and accurate than the affine-based solver.
****

Localized Concept Erasure for Text-to-Image Diffusion Models Using Training-Free Gated Low-Rank Adaptation

Byung Hyun Lee, Sungjin Lim, Se Young Chun; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18596-18606

Fine-tuning based concept erasing has demonstrated promising results in preventing generation of harmful contents from text-to-image diffusion models by removing target concepts while preserving remaining concepts. To maintain the generation capability of diffusion models after concept erasure, it is necessary to remove only the image region containing the target concept when it locally appears in an image, leaving other regions intact. However, prior arts often compromise fidelity of the other image regions in order to erase the localized target concept appearing in a specific area, thereby reducing the overall performance of image generation. To address these limitations, we first introduce a framework called localized concept erasure, which allows for the deletion of only the specific a

rea containing the target concept in the image while preserving the other region s. As a solution for the localized concept erasure, we propose a training-free a pproach, dubbed Gated Low-rank adaptation for Concept Erasure (GLoCE), that inje cts a lightweight module into the diffusion model. GLoCE consists of low-rank ma trices and a simple gate, determined only by several generation steps for concep ts without training. By directly applying GLoCE to image embeddings and designin g the gate to activate only for target concepts, GLoCE can selectively remove on ly the region of the target concepts, even when target and remaining concepts co exist within an image. Extensive experiments demonstrated GLoCE not only improve s the image fidelity to text prompts after erasing the localized target concepts , but also outperforms prior arts in efficacy, specificity, and robustness by la rge margin and can be extended to mass concept erasure.

****

PARC: A Quantitative Framework Uncovering the Symmetries within Vision Language Models

Jenny Schmalfuss, Nadine Chang, Vibashan VS, Maying Shen, Andres Bruhn, Jose M. Alvarez; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25081-25091

Vision language models (VLMs) respond to user-crafted text prompts and visual in puts, and are applied to numerous real-world problems. VLMs integrate visual mod alities with large language models (LLMs), which are well known to be prompt-sen sitive. Hence, it is crucial to determine whether VLMs inherit this instability to varying prompts. We therefore investigate which prompt variations VLMs are mo st sensitive to and which VLMs are most agnostic to prompt variations. To this e nd, we introduce PARC (Prompt Analysis via Reliability and Calibration), a VLM p rompt sensitivity analysis framework built on three pillars: (1) plausible promp t variations in both the language and vision domain, (2) a novel model reliabili ty score with built-in guarantees, and (3) a calibration step that enables datas et- and prompt-spanning prompt variation analysis. Regarding prompt variations, PARC's evaluation shows that VLMs mirror LLM language prompt sensitivity in the vision domain, and most destructive variations change the expected answer. Regar ding models, outstandingly robust VLMs among 22 evaluated models come from the I nternVL2 family. We further find indications that prompt sensitivity is linked t o training data. https://github.com/NVlabs/PARC

****

RoboTwin: Dual-Arm Robot Benchmark with Generative Digital Twins

Yao Mu, Tianxing Chen, Zanxin Chen, Shijia Peng, Zhiqian Lan, Zeyu Gao, Zhixuan Liang, Qiaojun Yu, Yude Zou, Mingkun Xu, Lunkai Lin, Zhiqiang Xie, Mingyu Ding, Ping Luo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27649-27660

In the rapidly advancing field of robotics, dual-arm coordination and complex ob ject manipulation are essential capabilities for developing advanced autonomous systems. However, the scarcity of diverse, high-quality demonstration data and r eal-world-aligned evaluation benchmarks severely limits such development. To add ress this, we introduce RoboTwin, a generative digital twin framework that uses 3D generative foundation models and large language models to produce diverse exp ert datasets and provide a real-world-aligned evaluation platform for dual-arm r obotic tasks. Specifically, RoboTwin creates varied digital twins of objects fro m single 2D images, generating realistic and interactive scenarios. It also intr oduces a spatial relation-aware code generation framework that combines object a nnotations with large language models to break down tasks, determine spatial con straints, and generate precise robotic movement code. Our framework offers a com prehensive benchmark with both simulated and real-world data, enabling standardi zed evaluation and better alignment between simulated training and real-world pe rformance. We validated our approach using the open-source COBOT Magic Robot pla tform. Policies pre-trained on RoboTwin-generated data and fine-tuned with limit ed real-world samples demonstrate significant potential for enhancing dual-arm r obotic manipulation systems by improving success rates by over 70% for single-ar m tasks and over 40% for dual-arm tasks compared to models trained solely on rea l-world data.

****
Population Normalization for Federated Learning
Zhuoyao Wang, Fan Yi, Peizhu Gong, Caitou He, Cheng Jin, Weizhong Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10214-10223
Batch normalization (BN) is widely recognized as an essential method in training deep neural networks, facilitating convergence and enhancing model stability. However, in Federated Learning (FL) contexts, where training data are typically heterogeneous and clients often face resource constraints, the effectiveness of BN is considerably limited for two primary reasons. First, the population statistics, specifically the mean and variance of these heterogeneous datasets, vary substantially, resulting in inconsistent BN layers across client models, which ultimately drives these models to diverge further. Second, estimating statistics from a mini-batch is often imprecise since the batch size has to be small in resource-limited clients. This paper introduces Population Normalization, a novel technique for FL, in which the statistics are learned as trainable parameters rather than calculated from mini-batches as in BN. Thus, our normalization layers are homogeneous among the clients and the adverse impact of small batch size is eliminated as the model can be well-trained even when the batch size equals to one. To enhance the flexibility of our method in practical applications, we investigate the role of stochastic uncertainty in BN's statistical estimation. When larger batch sizes are available, we demonstrate that injecting simple artificial noise can effectively mimic this stochastic uncertainty and improve the model's generalization capability. Experimental results validate the efficacy of our approach across various FL tasks.
****
AnimateAnything: Consistent and Controllable Animation for Video Generation
Guojun Lei, Chi Wang, Rong Zhang, Yikai Wang, Hong Li, Weiwei Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27946-27956
We propose a unified approach for video-controlled generation, enabling text-based guidance and manual annotations to control the generation of videos, similar to camera direction guidance. Specifically, we designed a two-stage algorithm. In the first stage, we convert all control information into frame-by-frame motion flows. In the second stage, we use these motion flows as guidance to control the final video generation. Additionally, to reduce instability in the generated videos caused by large motion variations (such as those from camera movement, object motion, or manual inputs), which can result in flickering or the intermittent disappearance of objects, we transform the temporal feature computation in the video model into frequency-domain feature computation. This is because frequency-domain signals better capture the essential characteristics of an image, and by ensuring consistency in the video's frequency-domain features, we can enhance temporal coherence and reduce flickering in the final generated video.
****
PRaDA: Projective Radial Distortion Averaging
Daniil Sinitsyn, Linus Härenstam-Nielsen, Daniel Cremers; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21902-21912
We tackle the problem of automatic calibration of radially distorted cameras in challenging conditions. Accurately determining distortion parameters typically requires either 1) solving the full Structure from Motion (SfM) problem involving camera poses, 3D points, and the distortion parameters, which is only possible if many images with sufficient overlap are provided, or 2) relying heavily on learning-based methods that are comparatively less accurate. In this work, we demonstrate that distortion calibration can be decoupled from 3D reconstruction, maintaining the accuracy of SfM-based methods while avoiding many of the associated complexities. This is achieved by working in Projective Space, where the geometry is unique up to a homography, which encapsulates all camera parameters except for distortion. Our proposed method, Projective Radial Distortion Averaging, averages multiple distortion estimates in a fully projective framework without creating 3d points and full bundle adjustment. By relying on pairwise projective re

lations, our methods support any feature-matching approaches without constructin g point tracks across multiple images.
****

GenAssets: Generating in-the-wild 3D Assets in Latent Space
Ze Yang, Jingkang Wang, Haowei Zhang, Sivabalan Manivasagam, Yun Chen, Raquel Ur tasun; Proceedings of the Computer Vision and Pattern Recognition Conference (CV PR), 2025, pp. 22392-22403
High-quality 3D assets for traffic participants are critical for multi-sensor si mulation, which is essential for the safe end-to-end development of autonomy. Bu ilding assets from in-the-wild data is key for diversity and realism, but existi ng neural-rendering based reconstruction methods are slow and generate assets th at render well only from viewpoints close to the original observations, limiting their usefulness in simulation. Recent diffusion-based generative models build complete and diverse assets, but perform poorly on in-the-wild driving scenes, w here observed actors are captured under sparse and limited fields of view, and a re partially occluded. In this work, we propose a 3D latent diffusion model that learns on in-the-wild LiDAR and camera data captured by a sensor platform and g enerates high-quality 3D assets with complete geometry and appearance. Key to ou r method is a "reconstruct-then-generate" approach that first leverages occlusio n-aware neural rendering trained over multiple scenes to build a high-quality la tent space for objects, and then trains a diffusion model that operates on the l atent space. We show our method outperforms existing reconstruction and generati on based methods, unlocking diverse and scalable content creation for simulation . Please visit https://waabi.ai/genassets for more details.
****

RipVIS: Rip Currents Video Instance Segmentation Benchmark for Beach Monitoring and Safety
Andrei Dumitriu, Florin Tatui, Florin Miron, Aakash Ralhan, Radu Tudor Ionescu, Radu Timofte; Proceedings of the Computer Vision and Pattern Recognition Confere nce (CVPR), 2025, pp. 3427-3437
Rip currents are strong, localized and narrow currents of water that flow outwar ds into the sea, causing numerous beach-related injuries and fatalities worldwid e. Accurate identification of rip currents remains challenging due to their amor phous nature and the lack of annotated data, which often requires expert knowled ge. To address these issues, we present RipVIS, a large-scale video instance seg mentation benchmark explicitly designed for rip current segmentation. RipVIS is an order of magnitude larger than previous datasets, featuring 184 videos (212,3 28 frames), of which 150 videos (163,528 frames) are with rip currents, collecte d from various sources, including drones, mobile phones, and fixed beach cameras . Our dataset encompasses diverse visual contexts, such as wave-breaking pattern s, sediment flows, and water color variations, across multiple global locations, including USA, Mexico, Costa Rica, Portugal, Italy, Greece, Romania, Sri Lanka, Australia and New Zealand. Most videos are annotated at 5 FPS to ensure accurac y in dynamic scenarios, supplemented by an additional 34 videos (48,800 frames) without rip currents. We conduct comprehensive experiments with Mask R-CNN, Casc ade Mask R-CNN, SparseInst and YOLO11, fine-tuning these models for the task of rip current segmentation. Results are reported in terms of multiple metrics, wit h a particular focus on the F2 score to prioritize recall and reduce false negat ives. To enhance segmentation performance, we introduce a novel post-processing step based on Temporal Confidence Aggregation (TCA). RipVIS aims to set a new st andard for rip current segmentation, contributing towards safer beach environmen ts. We offer a benchmark website to share data, models, and results with the res earch community, encouraging ongoing collaboration and future contributions, at https://ripvis.ai.
****

Low-Rank Adaptation in Multilinear Operator Networks for Security-Preserving Inc remental Learning
Huu Binh Ta, Duc Nguyen, Quyen Tran, Toan Tran, Tung Pham; Proceedings of the Co mputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24341-24350
In security-sensitive fields, data should be encrypted to protect against unauth

orized access and maintain confidentiality throughout processing. However, traditional networks like ViTs and CNNs return different results when processing original data versus its encrypted form, meaning that they require data to be decrypted, posing a security risk by exposing sensitive information. One solution for this issue is using polynomial networks, including state-of-the-art Multilinear Operator Networks, which return the same outputs given the real data and their encrypted forms under Leveled Fully Homomorphic Encryption. Nevertheless, these models are susceptible to catastrophic forgetting in incremental learning settings. Thus, this paper will present a new low-rank adaptation method combined with the Gradient Projection Memory mechanism to minimize the issue. Our proposal is compatible with Leveled Fully Homomorphic Encryption while achieving a sharp improvement in performance compared to existing models.
****

Camera Resection from Known Line Pencils and a Radially Distorted Scanline
Juan C. Dibene, Enrique Dunn; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15843-15851
We present a marker-based geometric estimation framework for the absolute pose of a camera by analyzing the 1D observations in a single radially distorted pixel scanline.We leverage a pair of known co-planar pencils of lines, along with lens distortion parameters, to propose an ensemble of solvers exploring the space of estimation strategies applicable to our setup.First, we present a minimal algebraic solver requiring only six measurements and yielding eight solutions, which relies on the intersection of two conics defined by one of the pencils of lines.Then, we present a unique closed-form geometric solver from seven measurements. Finally, we present an homography-based formulation amenable to linear least-squares from eight or more measurements.Our geometric framework constitutes a theoretical analysis on the minimum geometric context necessary to solve in closed form for the absolute pose of a single camera from a single radially distorted scanline.
****

ESCAPE: Equivariant Shape Completion via Anchor Point Encoding
Burak Bekci, Nassir Navab, Federico Tombari, Mahdi Saleh; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6480-6489
Shape completion, a crucial task in 3D computer vision, involves predicting and filling the missing regions of scanned or partially observed objects. Current methods expect known pose or canonical coordinates and do not perform well under varying rotations, limiting their real-world applicability. We introduce ESCAPE (Equivariant Shape Completion via Anchor Point Encoding), a novel framework designed to achieve rotation-equivariant shape completion. Our approach employs a distinctive encoding strategy by selecting anchor points from a shape and representing all points as a distance to all anchor points. This enables the model to capture a consistent, rotation-equivariant understanding of the object's geometry. ESCAPE leverages a transformer architecture to encode and decode the distance transformations, ensuring that generated shape completions remain accurate and equivariant under rotational transformations. Subsequently, we perform optimization to calculate the predicted shapes from the encodings. Experimental evaluations demonstrate that ESCAPE achieves robust, high-quality reconstructions across arbitrary rotations and translations, showcasing its effectiveness in real-world applications without additional pose estimation modules.
****

SPC-GS: Gaussian Splatting with Semantic-Prompt Consistency for Indoor Open-World Free-view Synthesis from Sparse Inputs
Guibiao Liao, Qing Li, Zhenyu Bao, Guoping Qiu, Kanglin Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11264-11274
3D Gaussian Splatting-based indoor open-world free-view synthesis approaches have shown significant performance with dense input images. However, they exhibit poor performance when confronted with sparse inputs, primarily due to the sparse distribution of Gaussian points and insufficient view supervision. To relieve these challenges, we propose SPC-GS, leveraging Scene-layout-based Gaussian Initialization (SGI) and Semantic-Prompt Consistency (SPC) Regularization for open-wor

ld free view synthesis with sparse inputs. Specifically, SGI provides a dense, s
cene-layout-based Gaussian distribution by utilizing view-changed images generat
ed from the video generation model and view-constraint Gaussian points densifica
tion. Additionally, SPC mitigates limited view supervision by employing semantic
-prompt-based consistency constraints developed by SAM2. This approach leverages
 available semantics from training views, serving as instructive prompts, to opt
imize visually overlapping regions in novel views with 2D and 3D consistency con
straints. Extensive experiments demonstrate the superior performance of SPC-GS a
cross Replica and ScanNet benchmarks. Notably, our SPC-GS achieves a 3.06 dB gai
n in PSNR for reconstruction quality and a 7.3% improvement in mIoU for open-wor
ld semantic segmentation.
****

M3amba: Memory Mamba is All You Need for Whole Slide Image Classification
Tingting Zheng, Kui Jiang, Yi Xiao, Sicheng Zhao, Hongxun Yao; Proceedings of th
e Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15601-156
10
Multi-instance learning (MIL) has demonstrated impressive performance in whole s
lide image (WSI) analysis. However, existing approaches struggle with undesirabl
e results and unbearable computational overhead due to the quadratic complexity
 of Transformers. Recently, Mamba has offered a feasible solution for modeling lo
ng-range dependencies with linear complexity. However, vanilla Mamba inherently
 suffers from contextual forgetting issues, making it ill-suited for capturing gl
obal dependencies across instances in large-scale WSIs. To address this, we prop
ose a memory-driven Mamba network, dubbed M3amba, to fully explore the global la
tent relations among instances. Specifically, M3amba retains and iteratively upd
ates historical information with a dynamic memory bank (DMB), thus overcoming th
e catastrophic forgetting defects of Mamba for long-term context representation.
 For better feature representation, M3amba involves an intra-group bidirectional
 Mamba (BiMamba) block to refine local interactions within groups. Meanwhile, we
 additionally perform cross-attention fusion to incorporate relevant historical
information across groups, facilitating richer inter-group connections. The join
t learning of inter- and intra-group representations with memory merits enables
M3amba with a more powerful capability for achieving accurate and comprehensive
WSI representation. Extensive experiments on four datasets demonstrate that M3am
ba outperforms the state-of-the-art by 6.2% and 7.0% in accuracy on the TCGA BRA
C and TCGA Lung datasets while maintaining low computational costs.
****

Satellite to GroundScape - Large-scale Consistent Ground View Generation from Sa
tellite Views
Ningli Xu, Rongjun Qin; Proceedings of the Computer Vision and Pattern Recogniti
on Conference (CVPR), 2025, pp. 6068-6077
Generating consistent ground-view images from satellite imagery is challenging,
primarily due to the large discrepancies in viewing angles and resolution betwee
n satellite and ground-level domains. Previous efforts mainly concentrated on si
ngle-view generation, often resulting in inconsistencies across neighboring grou
nd views. In this work, we propose a novel cross-view synthesis approach designe
d to overcome these challenges by ensuring consistency across ground-view images
 generated from satellite views. Our method, based on a fixed latent diffusion m
odel, introduces two conditioning modules: satellite-guided denoising, which ext
racts high-level scene layout to guide the denoising process, and satellite-temp
oral denoising, which captures camera motion to maintain consistency across mult
iple generated views. We further contribute a large-scale satellite-ground datas
et containing over 100,000 perspective pairs to facilitate extensive ground scen
e or video generation. Experimental results demonstrate that our approach outper
forms existing methods on perceptual and temporal metrics, achieving high photor
ealism and consistency in multi-view outputs. The project page is at https://gda
osu.github.io/sat2groundscape
****

Variance-Based Membership Inference Attacks Against Large-Scale Image Captioning
 Models

Daniel Samira, Edan Habler, Yuval Elovici, Asaf Shabtai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9210-9219
The proliferation of multi-modal generative models has introduced new privacy and security challenges, especially due to the risks of memorization and unintentional disclosure of sensitive information. This paper focuses on the vulnerability of multi-modal image captioning models to membership inference attacks (MIAs). These models, which synthesize textual descriptions from visual content, could inadvertently reveal personal or proprietary data embedded in their training datasets. We explore the feasibility of MIAs in the context of such models. Specifically, our approach leverages a variance-based strategy tailored for image captioning models, utilizing only image data without knowing the corresponding caption. We introduce the means-of-variance threshold attack (MVTA) and confidence-based weakly supervised attack (C-WSA) based on the metric, means-of-variance (MV), to assess variability among vector embeddings. Our experiments demonstrate that these models are susceptible to MIAs, indicating substantial privacy risks. The effectiveness of our methods is validated through rigorous evaluations on these real-world models, confirming the practical implications of our findings.
****

Redefining <Creative> in Dictionary: Towards an Enhanced Semantic Understanding of Creative Generation
Fu Feng, Yucheng Xie, Xu Yang, Jing Wang, Xin Geng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18444-18454
Creative remains an inherently abstract concept for both humans and diffusion models. While text-to-image (T2I) diffusion models can easily generate out-of-distribution concepts like "a blue banana", they struggle with generating combinatorial objects such as "a creative mixture that resembles a lettuce and a mantis", due to difficulties in understanding the semantic depth of "creative". Current methods rely heavily on synthesizing reference prompts or images to achieve a creative effect, typically requiring retraining for each unique creative output-a process that is computationally intensive and limits practical applications. To address this, we introduce CreTok, which brings meta-creativity to diffusion models by redefining "creative" as a new token, <CreTok>, thus enhancing models' semantic understanding for combinatorial creativity. CreTok achieves such redefinition by iteratively sampling diverse text pairs from our proposed CangJie dataset to form adaptive prompts and restrictive prompts, and then optimizing the similarity between their respective text embeddings. Extensive experiments demonstrate that <CreTok> enables the universal and direct generation of combinatorial creativity across diverse concepts without additional training, achieving state-of-the-art performance with improved text-image alignment and higher human preference ratings. Code will be made available at https://github.com/fu-feng/CreTok.
****

FiRe: Fixed-points of Restoration Priors for Solving Inverse Problems
Matthieu Terris, Ulugbek S. Kamilov, Thomas Moreau; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23185-23194
Selecting an appropriate prior to compensate for information loss due to the measurement operator is a fundamental challenge in imaging inverse problems. Implicit priors based on denoising neural networks have become central to widely-used frameworks such as Plug-and-Play (PnP) algorithms. In this work, we introduce Fixed-points of Restoration (FiRe) priors as a new framework for expanding the notion of priors in PnP to general restoration models beyond traditional denoising models. The key insight behind FiRe is that smooth images emerge as fixed points of the composition of a degradation operator with the corresponding restoration model. This enables us to derive an explicit formula for our implicit prior by quantifying invariance of images under this composite operation. Adopting this fixed-point perspective, we show how various restoration networks can effectively serve as priors for solving inverse problems. The FiRe framework further enables ensemble-like combinations of multiple restoration models as well as acquisition-informed restoration networks, all within a unified optimization approach. Experimental results validate the effectiveness of FiRe across various inverse problems, establishing a new paradigm for incorporating pretrained restoration mode

ls into PnP-like algorithms. Code available at https://github.com/matthieutrs/fire.

****

Learning Dynamic Collaborative Network for Semi-supervised 3D Vessel Segmentation

Jiao Xu, Xin Chen, Lihe Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10445-10454

In this paper, we present a new dynamic collaborative network for semi-supervised 3D vessel segmentation, termed DiCo. Conventional mean teacher (MT) methods typically employ a static approach, where the roles of the teacher and student models are fixed. However, due to the complexity of 3D vessel data, the teacher model may not always outperform the student model, leading to cognitive biases that can limit performance. To address this issue, we propose a dynamic collaborative network that allows the two models to dynamically switch their teacher-student roles. Additionally, we introduce a multi-view integration module to capture various perspectives of the inputs, mirroring the way doctors conduct medical analysis. We also incorporate adversarial supervision to constrain the shape of the segmented vessels in unlabeled data. In this process, the 3D volume is projected into 2D views to mitigate the impact of label inconsistencies. Experiments demonstrate that our DiCo method sets new state-of-the-art performance on three 3D vessel segmentation benchmarks. The code repository address is https://github.com/xujiaommcome/DiCo.

****

Temporal Alignment-Free Video Matching for Few-shot Action Recognition

SuBeen Lee, WonJun Moon, Hyun Seok Seong, Jae-Pil Heo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5412-5421

Few-Shot Action Recognition (FSAR) aims to train a model with only a few labeled video instances. A key challenge in FSAR is handling divergent narrative trajectories for precise video matching. While the frame- and tuple-level alignment approaches have been promising, their methods heavily rely on pre-defined and length-dependent alignment units (e.g., frames or tuples), which limits flexibility for actions of varying lengths and speeds. In this work, we introduce a novel TEmporal Alignment-free Matching (TEAM) approach, which eliminates the need for temporal units in action representation and brute-force alignment during matching. Specifically, TEAM represents each video with a fixed set of pattern tokens that capture globally discriminative clues within the video instance regardless of action length or speed, ensuring its flexibility. Furthermore, TEAM is inherently efficient, using token-wise comparisons to measure similarity between videos, unlike existing methods that rely on pairwise comparisons for temporal alignment. Additionally, we propose an adaptation process that identifies and removes common information across classes, establishing clear boundaries even between novel categories. Extensive experiments demonstrate the effectiveness of TEAM. Codes are available at github.com/leesb7426/TEAM.

****

OSLoPrompt: Bridging Low-Supervision Challenges and Open-Set Domain Generalization in CLIP

Mohamad Hassan N C, Divyam Gupta, Mainak Singha, Sai Bhargav Rongali, Ankit Jha, Muhammad Haris Khan, Biplab Banerjee; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10110-10120

We introduce Low-Shot Open-Set Domain Generalization (LSOSDG), a novel paradigm unifying low-shot learning with open-set domain generalization (ODG). While prompt-based methods using models like CLIP have advanced DG, they falter in low-data regimes (e.g., 1-shot) and lack precision in detecting open-set samples with fine-grained semantics related to training classes. To address these challenges, we propose OSLoPrompt, an advanced prompt-learning framework for CLIP with two core innovations. First, to manage limited supervision across source domains and improve DG, we introduce a domain-agnostic prompt-learning mechanism that integrates adaptable domain-specific cues and visually guided semantic attributes through a novel cross-attention module, besides being supported by learnable domain- and class-generic visual prompts to enhance cross-modal adaptability. Second, t

o improve outlier rejection during inference, we classify unfamiliar samples as "unknown" and train specialized prompts with systematically synthesized pseudo-open samples that maintain fine-grained relationships to known classes, generated through a targeted query strategy with off-the-shelf foundation models. This strategy enhances feature learning, enabling our model to detect open samples with varied granularity more effectively. Extensive evaluations across five benchmarks demonstrate that OSLoPrompt establishes a new state-of-the-art in LSOSDG, significantly outperforming existing methods.
****
Dinomaly: The Less Is More Philosophy in Multi-Class Unsupervised Anomaly Detection
Jia Guo, Shuai Lu, Weihang Zhang, Fang Chen, Huiqi Li, Hongen Liao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20405-20415
Recent studies highlighted a practical setting of unsupervised anomaly detection (UAD) that builds a unified model for multi-class images. Despite various advancements addressing this challenging task, the detection performance under the multi-class setting still lags far behind state-of-the-art class-separated models. Our research aims to bridge this substantial performance gap. In this paper, we present Dinomaly, a minimalist reconstruction-based anomaly detection framework that harnesses pure Transformer architectures without relying on complex designs, additional modules, or specialized tricks. Given this powerful framework consisting of only Attentions and MLPs, we found four simple components that are essential to multi-class anomaly detection: (1) Scalable foundation Transformers that extracts universal and discriminative features, (2) Noisy Bottleneck where pre-existing Dropouts do all the noise injection tricks, (3) Linear Attention that naturally cannot focus, and (4) Loose Reconstruction that does not force layer-to-layer and point-by-point reconstruction. Extensive experiments are conducted across popular anomaly detection benchmarks including MVTec-AD, VisA, and Real-IAD. Our proposed Dinomaly achieves impressive image-level AUROC of 99.6%, 98.7%, and 89.3% on the three datasets respectively, which is not only superior to state-of-the-art multi-class UAD methods, but also achieves the most advanced class-separated UAD records.
****
VLog: Video-Language Models by Generative Retrieval of Narration Vocabulary
Kevin Qinghong Lin, Mike Zheng Shou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3218-3228
Human daily activities can be concisely narrated as sequences of routine events (e.g., turning off an alarm) in video streams, forming an event vocabulary. Motivated by this, we introduce **VLog**, a novel video understanding framework that defines video narrations as a vocabulary, going beyond the typical subword vocabularies in existing generative video-language models. Built on the lightweight language model GPT-2, **VLog** features three key innovations:1. **A Generative Retrieval Model** Marrying the language model's complex reasoning capabilities with contrastive retrieval's efficient similarity search.2. **A Hierarchical Vocabulary** Derived from large-scale video narrations using our narration pair encoding algorithm, enabling efficient indexing of specific events (e.g., cutting a tomato) by identifying broader scenarios (e.g., kitchen) with expressive postfixes (e.g., by the left hand).3. **A Vocabulary Update Strategy** Leveraging generative models to extend the vocabulary for novel events encountered during inference.To validate our approach, we introduce **VidCab-Eval**, a development set requiring concise narrations with reasoning relationships (e.g., before and after). Experiments on **EgoSchema**, **COIN**, and **HiREST** further demonstrate the effectiveness of **VLog**, highlighting its ability to generate concise, contextually accurate, and efficient narrations. This offers a novel perspective on video understanding.
****
CoMBO: Conflict Mitigation via Branched Optimization for Class Incremental Segmentation
Kai Fang, Anqi Zhang, Guangyu Gao, Jianbo Jiao, Chi Harold Liu, Yunchao Wei; Pro

ceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25667-25676
Effective Class Incremental Segmentation (CIS) requires simultaneously mitigating catastrophic forgetting and ensuring sufficient plasticity to integrate new classes. The inherent conflict above often leads to a back-and-forth, which turns the objective into finding the balance between the performance of previous (old) and incremental (new) classes. To address this conflict, we introduce a novel approach, Conflict Mitigation via Branched Optimization (CoMBO). Within this approach, we present the Query Conflict Reduction module, designed to explicitly refine queries for new classes through lightweight, class-specific adapters. This module provides an additional branch for the acquisition of new classes while preserving the original queries for distillation. Moreover, we develop two strategies to further mitigate the conflict following the branched structure, i.e., the Half-Learning Half-Distillation (HDHL) over classification probabilities, and the Importance-Based Knowledge Distillation (IKD) over query features. HDHL selectively engages in learning for classification probabilities of queries that match the ground truth of new classes, while aligning unmatched ones to the corresponding old probabilities, thus ensuring retention of old knowledge while absorbing new classes via learning negative samples. Meanwhile, IKD assesses the importance of queries based on their matching degree to old classes, prioritizing the distillation of important features and allowing less critical features to evolve. Extensive experiments in Class Incremental Panoptic and Semantic Segmentation settings have demonstrated the superior performance of CoMBO. Project page: https://guangyu-ryan.github.io/CoMBO.
****

Forensics-Bench: A Comprehensive Forgery Detection Benchmark Suite for Large Vision Language Models
Jin Wang, Chenghui Lv, Xian Li, Shichao Dong, Huadong Li, Kelu Yao, Chao Li, Wenqi Shao, Ping Luo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4233-4245
Recently, the rapid development of AIGC has significantly boosted the diversities of fake media spread in the Internet, posing unprecedented threats to social security, politics, law, and etc.To detect the ever-increasingly **diverse** malicious fake media in the new era of AIGC, recent studies have proposed to exploit Large Vision Language Models (LVLMs) to design **robust** forgery detectors due to their impressive performance on a **wide** range of multimodal tasks.However, it still lacks a comprehensive benchmark designed to comprehensively assess LVLMs' discerning capabilities on forgery media.To fill this gap, we present Forensics-Bench, a new forgery detection evaluation benchmark suite to assess LVLMs across massive forgery detection tasks, requiring comprehensive recognition, location and reasoning capabilities on diverse forgeries.Forensics-Bench comprises 63,292 meticulously curated multi-choicevisual questions, covering 112 unique forgery detection types from 5 perspectives: forgery semantics, forgery modalities, forgery tasks, forgery types and forgery models.We conduct thorough evaluations on 22 open-sourced LVLMs and 3 proprietary models GPT-4o, Gemini 1.5 Pro, and Claude 3.5 Sonnet, highlighting the significant challenges of comprehensive forgery detection posed by Forensics-Bench.We anticipate that Forensics-Bench will motivate the community to advance the frontier of LVLMs, striving for all-around forgery detectors in the era of AIGC.
****

Detect-and-Guide: Self-regulation of Diffusion Models for Safe Text-to-Image Generation via Guideline Token Optimization
Feifei Li, Mi Zhang, Yiming Sun, Min Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13252-13262
Text-to-image diffusion models have achieved state-of-the-art results in synthesis tasks; however, there is a growing concern about their potential misuse in creating harmful content. To mitigate these risks, post-hoc model intervention techniques, such as concept unlearning and safety guidance, have been developed. However, fine-tuning model weights or adapting the hidden states of the diffusion model operates in an uninterpretable way, making it unclear which part of the in

termediate variables is responsible for unsafe generation. These interventions s everely affect the sampling trajectory when erasing harmful concepts from comple x, multi-concept prompts, thus hindering their practical use in real-world setti ngs. In this work, we propose the safe generation framework Detect-and-Guide (DA G), leveraging the internal knowledge of diffusion models to perform self-diagno sis and fine-grained self-regulation during the sampling process. DAG first dete cts harmful concepts from noisy latents using refined cross-attention maps of op timized tokens, then applies safety guidance with adaptive strength and editing regions to negate unsafe generation. The optimization only requires a small anno tated dataset and can provide precise detection maps with generalizability and c oncept specificity. Moreover, DAG does not require fine-tuning of diffusion mode ls, and therefore introduces no loss to their generation diversity. Experiments on erasing sexual content show that DAG achieves state-of-the-art safe generatio n performance, balancing harmfulness mitigation and text-following performance o n multi-concept real-world prompts.

****

MirrorVerse: Pushing Diffusion Models to Realistically Reflect the World

Ankit Dhiman, Manan Shah, R Venkatesh Babu; Proceedings of the Computer Vision a nd Pattern Recognition Conference (CVPR), 2025, pp. 11239-11249

Diffusion models have become central to various image editing tasks, yet they of ten fail to fully adhere to physical laws, particularly with effects like shadow s, reflections, and occlusions. In this work, we address the challenge of genera ting photorealistic mirror reflections using diffusion-based generative models. Despite extensive training data, existing diffusion models frequently overlook t he nuanced details crucial to authentic mirror reflections. Recent approaches ha ve attempted to resolve this by creating synthetic datasets and framing reflecti on generation as an inpainting task; however, they struggle to generalize across different object orientations and positions relative to the mirror. Our method overcomes these limitations by introducing key augmentations into the synthetic data pipeline: (1) random object positioning, (2) randomized rotations, and (3) grounding of objects, significantly enhancing generalization across poses and pl acements. To further address spatial relationships and occlusions in scenes with multiple objects, we implement a strategy to pair objects during dataset genera tion, resulting in a dataset robust enough to handle these complex scenarios. Ac hieving generalization to real-world scenes remains a challenge, so we introduce a three-stage training curriculum to develop the MirrorFusion 2.0 model, aimed at improving real-world performance. We provide extensive qualitative and quanti tative evaluations to support our approach. The project page is available at: ht tps://mirror-verse.github.io/.

****

EAP-GS: Efficient Augmentation of Pointcloud for 3D Gaussian Splatting in Few-sh ot Scene Reconstruction

Dongrui Dai, Yuxiang Xing; Proceedings of the Computer Vision and Pattern Recogn ition Conference (CVPR), 2025, pp. 16498-16507

3D Gaussian splatting (3DGS) has shown impressive performance in 3D scene recons truction. However, it suffers from severe degradation when the number of trainin g views is limited, resulting in blur and floaters. Many works have been devoted to standardize the optimization process of 3DGS through regularization techniqu es. However, we identify that inadequate initialization is a critical issue over looked by current studies. To address this, we propose EAP-GS, a method to enhan ce initialization for fast, accurate, and stable few-shot scene reconstruction. Specifically, we introduce an Attentional Pointcloud Augmentation (APA) techniqu e, which retains two-view tracks as an option for pointcloud generation. Additio nally, the scene complexity is used to determine the required density distributi on, thereby constructing a better pointcloud. We implemented APA by extending St ructure-From-Motion (SFM) to focus on pointcloud generation in regions with comp lex structure but sparse pointcloud distribution, which significantly increases the number of valuable points and effectively harmonizes the density distributio n. A better pointcloud leads to more accurate scene geometry and mitigates local overfitting during reconstruction stage. Furthermore, our APA can be framed as

a modular augmentation to existing methods with minimal overhead. Experimental r
esults from various indoor and outdoor scenes demonstrate that the proposed EAP-
GS achieves outstanding scene reconstruction performance and surpasses state-of-
the-art methods. Project page: https://osierddr.github.io/eapgs/
****

Empowering Large Language Models with 3D Situation Awareness
Zhihao Yuan, Yibo Peng, Jinke Ren, Yinghong Liao, Yatong Han, Chun-Mei Feng, Hen
gshuang Zhao, Guanbin Li, Shuguang Cui, Zhen Li; Proceedings of the Computer Vis
ion and Pattern Recognition Conference (CVPR), 2025, pp. 19435-19445
Driven by the great success of Large Language Models (LLMs) in the 2D image doma
in, their applications in 3D scene understanding has emerged as a new trend. A k
ey difference between 3D and 2D is that the situation of an egocentric observer
in 3D scenes can change, resulting in different descriptions (e.g., "left" or "r
ight"). However, current LLM-based methods overlook the egocentric perspective a
nd simply use datasets from a global viewpoint. To address this issue, we propos
e a novel approach to automatically generate a situation-aware dataset by levera
ging the scanning trajectory during data collection and utilizing Vision-Languag
e Models (VLMs) to produce high-quality captions and question-answer pairs. Furt
hermore, we introduce a situation grounding module to explicitly predict the pos
ition and orientation of observer's viewpoint, thereby enabling LLMs to ground s
ituation description in 3D scenes. We evaluate our approach on several benchmark
s, demonstrating that our method effectively enhances the 3D situational awarene
ss of LLMs while significantly expanding existing datasets and reducing manual e
ffort.
****

Forensic Self-Descriptions Are All You Need for Zero-Shot Detection, Open-Set So
urce Attribution, and Clustering of AI-generated Images
Tai D. Nguyen, Aref Azizpour, Matthew C. Stamm; Proceedings of the Computer Visi
on and Pattern Recognition Conference (CVPR), 2025, pp. 3040-3050
The emergence of advanced AI-based tools to generate realistic images poses sign
ificant challenges for forensic detection and source attribution, especially as
new generative techniques appear rapidly. Traditional methods often fail to gene
ralize to unseen generators due to reliance on features specific to known source
s during training. To address this problem, we propose a novel approach that exp
licitly models forensic microstructures--subtle, pixel-level patterns unique to
the image creation process. Using only real images in a self-supervised manner,
we learn a set of diverse predictive filters to extract residuals that capture d
ifferent aspects of these microstructures. By jointly modeling these residuals a
cross multiple scales, we obtain a compact model whose parameters constitute a u
nique forensic self-description for each image. This self-description enables us
 to perform zero-shot detection of synthetic images, open-set source attribution
 of images, and clustering based on source without prior knowledge. Extensive ex
periments demonstrate that our method achieves superior accuracy and adaptabilit
y compared to competing techniques, advancing the state of the art in synthetic
media forensics.
****

EchoTraffic: Enhancing Traffic Anomaly Understanding with Audio-Visual Insights
Zhenghao Xing, Hao Chen, Binzhu Xie, Jiaqi Xu, Ziyu Guo, Xuemiao Xu, Jianye Hao,
 Chi-Wing Fu, Xiaowei Hu, Pheng-Ann Heng; Proceedings of the Computer Vision and
 Pattern Recognition Conference (CVPR), 2025, pp. 19098-19108
Traffic Anomaly Understanding (TAU) is essential for improving public safety and
 transportation efficiency by enabling timely detection and response to incident
s. Beyond existing methods, which rely largely on visual data, we propose to con
sider audio cues, a valuable source that offers strong hints to anomaly scenario
s such as crashes and honking. Our contributions are twofold. First, we compile
AV-TAU, the first large-scale audio-visual dataset for TAU, providing 29,865 tra
ffic anomaly videos and 149,325 Q&A pairs, while supporting five essential TAU t
asks. Second, we develop EchoTraffic, a multimodal LLM that integrates audio and
 visual data for TAU, through our audio-insight frame selector and dynamic conne
ctor to effectively extract crucial audio cues for anomaly understanding with a

two-phase training framework. Experimental results on AV-TAU manifest that EchoTraffic sets a new SOTA performance in TAU, outperforming the existing multimodal LLMs. Our contributions, including AV-TAU and EchoTraffic, pave a new direction for multimodal TAU.
****

FlexDrive: Toward Trajectory Flexibility in Driving Scene Gaussian Splatting Reconstruction and Rendering
Jingqiu Zhou, Lue Fan, Linjiang Huang, Xiaoyu Shi, Si Liu, Zhaoxiang Zhang, Hongsheng Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1549-1558
Driving scene reconstruction and rendering have advanced significantly using the 3D Gaussian Splatting.However, most prior research has focused on the rendering quality along a pre-recorded vehicle path and struggles to generalize to out-of-path viewpoints, which is caused by the lack of high-quality supervision in those out-of-path views. To address this issue, we introduce an Inverse View Warping technique to create compact and high-quality images as supervision for the reconstruction of the out-of-path views, enabling high-quality rendering results for those views.For accurate and robust inverse view warping, a depth bootstrap strategy is proposed to obtain on-the-fly dense depth maps during the optimization process, overcoming the sparsity and incompleteness of LiDAR depth data.Our method achieves superior in-path and out-of-path reconstruction and rendering performance on the widely used Waymo Open dataset.In addition, a simulator-based benchmark is proposed to obtain the out-of-path ground truth and quantitatively evaluate the performance of out-of-path rendering, where our method outperforms previous methods by a significant margin.
****

Taming Video Diffusion Prior with Scene-Grounding Guidance for 3D Gaussian Splatting from Sparse Inputs
Yingji Zhong, Zhihao Li, Dave Zhenyu Chen, Lanqing Hong, Dan Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6133-6143
Despite recent successes in novel view synthesis using 3D Gaussian Splatting (3DGS), modeling scenes with sparse inputs remains a challenge. In this work, we address two critical yet overlooked issues in real-world sparse-input modeling: extrapolation and occlusion. To tackle these issues, we propose to use a reconstruction by generation pipeline that leverages learned priors from video diffusion models to provide plausible interpretations for regions outside the field of view or occluded. However, the generated sequences exhibit inconsistencies that do not fully benefit subsequent 3DGS modeling. To address the challenge of inconsistency, we introduce a novel scene-grounding guidance based on rendered sequences from an optimized 3DGS, which tames the diffusion model to generate consistent sequences. This guidance is training-free and does not require any fine-tuning of the diffusion model. To facilitate holistic scene modeling, we also propose a trajectory initialization method. It effectively identifies regions that are outside the field of view and occluded. We further design a scheme tailored for 3DGS optimization with generated sequences. Experiments demonstrate that our method significantly improves upon the baseline and achieves state-of-the-art performance on challenging benchmarks.
****

Interactive Medical Image Segmentation: A Benchmark Dataset and Baseline
Junlong Cheng, Bin Fu, Jin Ye, Guoan Wang, Tianbin Li, Haoyu Wang, Ruoyu Li, He Yao, Junren Cheng, Jingwen Li, Yanzhou Su, Min Zhu, Junjun He; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20841-20851
Interactive Medical Image Segmentation (IMIS) has long been constrained by the limited availability of large-scale, diverse, and densely annotated datasets, which hinders model generalization and consistent evaluation across different models. In this paper, we introduce the IMed-361M benchmark dataset, a significant advancement in general IMIS research. First, we collect and standardize over 6.4 million medical images and their corresponding ground truth masks from multiple d

ata sources. Then, leveraging the strong object recognition capabilities of a vi sion foundational model, we automatically generated dense interactive masks for each image and ensured their quality through rigorous quality control and granul arity management. Unlike previous datasets, which are limited by specific modali ties or sparse annotations, IMed-361M spans 14 modalities and 204 segmentation t argets, totaling 361 million masks--an average of 56 masks per image. Finally, w e developed an IMIS baseline network on this dataset that supports high-quality mask generation through interactive inputs, including clicks, bounding boxes, te xt prompts, and their combinations. We evaluate its performance on medical image segmentation tasks from multiple perspectives, demonstrating superior accuracy and scalability compared to existing interactive segmentation models. To facilit ate research on foundational models in medical computer vision, we release the I Med-361M and model at https://github.com/uni-medical/IMIS-Bench.

****

GigaHands: A Massive Annotated Dataset of Bimanual Hand Activities

Rao Fu, Dingxi Zhang, Alex Jiang, Wanjia Fu, Austin Funk, Daniel Ritchie, Srinat h Sridhar; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17461-17474

Understanding bimanual human hand activities is a critical problem in AI and rob otics. We cannot build large models of bimanual activities because existing data sets lack the scale, coverage of diverse hand activities, and detailed annotatio ns. We introduce GigaHands, a massive annotated dataset capturing 34 hours of bi manual hand activities from 56 subjects and 417 objects, totaling 14k motion cli ps derived from 183 million frames paired with 84k text annotations. Our markerl ess capture setup and data acquisition protocol enable fully automatic 3D hand a nd object estimation while minimizing the effort required for text annotation. T he scale and diversity of GigaHands enable broad applications, including text-dr iven action synthesis, hand motion captioning, and dynamic radiance field recons truction.

****

AutoSSVH: Exploring Automated Frame Sampling for Efficient Self-Supervised Video Hashing

Niu Lian, Jun Li, Jinpeng Wang, Ruisheng Luo, Yaowei Wang, Shu-Tao Xia, Bin Chen ; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18881-18890

Self-Supervised Video Hashing (SSVH) compresses videos into hash codes for effic ient indexing and retrieval using unlabeled training videos. Existing approaches rely on random frame sampling to learn video features and treat all frames equa lly. This results in suboptimal hash codes, as it ignores frame-specific informa tion density and reconstruction difficulty. To address this limitation, we propo se a new framework, termed AutoSSVH, that employs adversarial frame sampling wit h hash-based contrastive learning. Our adversarial sampling strategy automatical ly identifies and selects challenging frames with richer information for reconst ruction, enhancing encoding capability. Additionally, we introduce a hash compon ent voting strategy and a point-to-set (P2Set) hash-based contrastive objective, which help capture complex inter-video semantic relationships in the Hamming sp ace and improve the discriminability of learned hash codes. Extensive experiment s demonstrate that AutoSSVH achieves superior retrieval efficacy and efficiency compared to state-of-the-art approaches. Code is available at https://github.com /EliSpectre/CVPR25-AutoSSVH.

****

Augmenting Multimodal LLMs with Self-Reflective Tokens for Knowledge-based Visua l Question Answering

Federico Cocchi, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, Rita Cucc hiara; Proceedings of the Computer Vision and Pattern Recognition Conference (CV PR), 2025, pp. 9199-9209

Multimodal LLMs (MLLMs) are the natural extension of large language models to ha ndle multimodal inputs, combining text and image data. They have recently garner ed attention due to their capability to address complex tasks involving both mod alities. However, their effectiveness is limited to the knowledge acquired durin

g training, which restricts their practical utility. In this work, we introduce a novel method to enhance the adaptability of MLLMs by integrating external knowledge sources. Our proposed model, Reflective LLaVA (ReflectiVA), utilizes reflective tokens to dynamically determine the need for external knowledge and predict the relevance of information retrieved from an external database. Tokens are trained following a two-stage two-model training recipe. This ultimately enables the MLLM to manage external knowledge while preserving fluency and performance on tasks where external knowledge is not needed. Through our experiments, we demonstrate the efficacy of ReflectiVA for knowledge-based visual question answering, highlighting its superior performance compared to existing methods. Source code and trained models are publicly available at https://aimagelab.github.io/ReflectiVA.

****

DifIISR: A Diffusion Model with Gradient Guidance for Infrared Image Super-Resolution

Xingyuan Li, Zirui Wang, Yang Zou, Zhixin Chen, Jun Ma, Zhiying Jiang, Long Ma, Jinyuan Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7534-7544

Infrared imaging is essential for autonomous driving and robotic operations as a supportive modality due to its reliable performance in challenging environments. Despite its popularity, the limitations of infrared cameras, such as low spatial resolution and complex degradations, consistently challenge imaging quality and subsequent visual tasks. Hence, infrared image super-resolution (IISR) has been developed to address this challenge. While recent developments in diffusion models have greatly advanced this field, current methods to solve it either ignore the unique modal characteristics of infrared imaging or overlook the machine perception requirements. To bridge these gaps, we propose DifIISR, an infrared image super-resolution diffusion model optimized for visual quality and perceptual performance. Our approach achieves task-based guidance for diffusion by injecting gradients derived from visual and perceptual priors into the noise during the reverse process. Specifically, we introduce an infrared thermal spectrum distribution regulation to preserve visual fidelity, ensuring that the reconstructed infrared images closely align with high-resolution images by matching their frequency components. Subsequently, we incorporate various visual foundational models as the perceptual guidance for downstream visual tasks, infusing generalizable perceptual features beneficial for detection and segmentation. As a result, our approach gains superior visual results while attaining State-Of-The-Art downstream task performance.

****

Recurrent Feature Mining and Keypoint Mixup Padding for Category-Agnostic Pose Estimation

Junjie Chen, Weilong Chen, Yifan Zuo, Yuming Fang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22035-22044

Category-agnostic pose estimation aims to locate keypoints on query images according to a few annotated support images for arbitrary novel classes. Existing methods generally extract support features via heatmap pooling, and obtain interacted features from support and query via cross-attention. Hence, these works neglect to mine fine-grained and structure-aware (FGSA) features from both support and query images, which are crucial for pixel-level keypoint localization. To this end, we propose a novel yet concise framework, which recurrently mines FGSA features from both support and query images. Specifically, we design a FGSA mining module based on deformable attention mechanism. On the one hand, we mine fine-grained features by applying deformable attention head over multi-scale feature maps. On the other hand, we mine structure-aware features by offsetting the reference points of keypoints to their linked keypoints. By means of above module, we recurrently mine FGSA features from support and query images, and thus obtain better support features and query estimations. In addition, we propose to use mixup keypoints to pad various classes to a unified keypoint number, which could provide richer supervision than the zero padding used in existing works. We conduct extensive experiments and in-depth studies on large-scale MP-100 dataset, and o

utperform SOTA method dramatically (+3.2% PCK\@0.05).
****
FrugalNeRF: Fast Convergence for Extreme Few-shot Novel View Synthesis without L
earned Priors
Chin-Yang Lin, Chung-Ho Wu, Chang-Han Yeh, Shih-Han Yen, Cheng Sun, Yu-Lun Liu;
Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 20
25, pp. 11227-11238
Neural Radiance Fields (NeRF) face significant challenges in extreme few-shot sc
enarios, primarily due to overfitting and long training times. Existing methods,
 such as FreeNeRF and SparseNeRF, use frequency regularization or pre-trained pr
iors but struggle with complex scheduling and bias. We introduce FrugalNeRF, a n
ovel few-shot NeRF framework that leverages weight-sharing voxels across multipl
e scales to efficiently represent scene details. Our key contribution is a cross
-scale geometric adaptation scheme that selects pseudo ground truth depth based
on reprojection errors across scales. This guides training without relying on ex
ternally learned priors, enabling full utilization of the training data. It can
also integrate pre-trained priors, enhancing quality without slowing convergence
. Experiments on LLFF, DTU, and RealEstate-10K show that FrugalNeRF outperforms
 other few-shot NeRF methods while significantly reducing training time, making
it a practical solution for efficient and accurate 3D scene reconstruction.
****
3D-GSW: 3D Gaussian Splatting for Robust Watermarking
Youngdong Jang, Hyunje Park, Feng Yang, Heeju Ko, Euijin Choo, Sangpil Kim; Proc
eedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025,
pp. 5938-5948
As 3D Gaussian Splatting (3D-GS) gains significant attention and its commercial
usage increases, the need for watermarking technologies to prevent unauthorized
use of the 3D-GS models and rendered images has become increasingly important. I
n this paper, we introduce a robust watermarking method for 3D-GS that secures c
opyright of both the model and its rendered images. Our proposed method remains
robust against distortions in rendered images and model attacks while maintainin
g high rendering quality. To achieve these objectives, we present Frequency-Guid
ed Densification (FGD), which removes 3D Gaussians based on their contribution t
o rendering quality, enhancing real-time rendering and the robustness of the mes
sage. FGD utilizes Discrete Fourier Transform to split 3D Gaussians in high-freq
uency areas, improving rendering quality. Furthermore, we employ a gradient mask
 for 3D Gaussians and design a wavelet-subband loss to enhance rendering quality
. Our experiments show that our method embeds the message in the rendered images
 invisibly and robustly against various attacks, including model distortion. Our
 method achieves superior performance in both rendering quality and watermark ro
bustness while improving real-time rendering efficiency. Project page: https://k
uai-lab.github.io/cvpr20253dgsw/
****
Pioneering 4-Bit FP Quantization for Diffusion Models: Mixup-Sign Quantization a
nd Timestep-Aware Fine-Tuning
Maosen Zhao, Pengtao Chen, Chong Yu, Yan Wen, Xudong Tan, Tao Chen; Proceedings
of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1813
4-18143
Model quantization reduces the bit-width of weights and activations, improving m
emory efficiency and inference speed in diffusion models. However, achieving 4-b
it quantization remains challenging. Existing methods, primarily based on intege
r quantization and post-training quantization fine-tuning, struggle with inconsi
stent performance. Inspired by the success of floating-point (FP) quantization i
n large language models, we explore low-bit FP quantization for diffusion models
 and identify key challenges: the failure of signed FP quantization to handle as
ymmetric activation distributions, the insufficient consideration of temporal co
mplexity in the denoising process during fine-tuning, and the misalignment betwe
en fine-tuning loss and quantization error. To address these challenges, we prop
ose the mixup-sign floating-point quantization (MSFP) framework, first introduci
ng unsigned FP quantization in model quantization, along with timestep-aware LoR

A (TALoRA) and denoising-factor loss alignment (DFA), which ensure precise and s
table fine-tuning. Extensive experiments show that we are the first to achieve s
uperior performance in 4-bit FP quantization for diffusion models, outperforming
 existing PTQ fine-tuning methods in 4-bit INT quantization.
****

OpenING: A Comprehensive Benchmark for Judging Open-ended Interleaved Image-Text
 Generation
Pengfei Zhou, Xiaopeng Peng, Jiajun Song, Chuanhao Li, Zhaopan Xu, Yue Yang, Ziy
ao Guo, Hao Zhang, Yuqi Lin, Yefei He, Lirui Zhao, Shuo Liu, Tianhua Li, Yuxuan
Xie, Xiaojun Chang, Yu Qiao, Wenqi Shao, Kaipeng Zhang; Proceedings of the Compu
ter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 56-66
Multimodal Large Language Models (MLLMs) have made significant strides in visual
 understanding and generation tasks. However, generating interleaved image-text
content remains a challenge, which requires integrated multimodal understanding
and generation abilities. While the progress in unified models offers new soluti
ons, existing benchmarks are insufficient for evaluating these methods due to da
ta size and diversity limitations. To bridge this gap, we introduce OpenING, a c
omprehensive benchmark comprising 5,400 high-quality human-annotated instances a
cross 56 real-world tasks. OpenING covers diverse daily scenarios such as travel
 guide, design, and brainstorming, offering a robust platform for challenging in
terleaved generation methods. In addition, we present IntJudge, a judge model fo
r evaluating open-ended multimodal generation methods. Trained with a novel data
 pipeline, our IntJudge achieves an agreement rate of 82.42% with human judgment
s, outperforming GPT-based evaluators by 11.34%. Extensive experiments on OpenIN
G reveal that current interleaved generation methods still have substantial room
 for improvement. Key findings on interleaved image-text generation are further
presented to guide the development of next-generation models.
****

Dual Exposure Stereo for Extended Dynamic Range 3D Imaging
Juhyung Choi, Jinnyeong Kim, Seokjun Choi, Jinwoo Lee, Samuel Brucker, Mario Bij
elic, Felix Heide, Seung-Hwan Baek; Proceedings of the Computer Vision and Patte
rn Recognition Conference (CVPR), 2025, pp. 6283-6293
Achieving robust stereo 3D imaging under diverse illumination conditions is an i
mportant however challenging task, largely due to the limited dynamic ranges (DR
s) of cameras, which are significantly smaller than real world DR. As a result,
the accuracy of existing stereo depth estimation methods is often compromised by
 under- or over-exposed images.In this work, we introduce dual-exposure stereo f
or extended dynamic range 3D imaging. We develop automatic dual-exposure control
 method that adjusts the dual exposures, diverging them when the scene DR exceed
s the camera DR, thereby providing information about broader DR. From the captur
ed dual-exposure stereo images, we estimate depth by developing a motion-aware d
ual-exposure stereo depth network.To validate our proposed method, we develop a
robot-vision system, collect real-world stereo video datasets, and generate a sy
nthetic dataset. Our approach outperforms traditional exposure control and depth
 estimation methods.
****

Unbiasing through Textual Descriptions: Mitigating Representation Bias in Video
Benchmarks
Nina Shvetsova, Arsha Nagrani, Bernt Schiele, Hilde Kuehne, Christian Rupprecht;
 Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2
025, pp. 29050-29059
We propose a new "Unbiased through Textual Description (UTD)" video benchmark ba
sed on unbiased subsets of existing video classification and retrieval datasets
to enable a more robust assessment of video understanding capabilities. Namely,
we tackle the problem that current video benchmarks may suffer from different re
presentation biases, e.g., object bias or single-frame bias, where mere recognit
ion of objects or utilization of only a single frame is sufficient for correct p
rediction. We leverage VLMs and LLMs to analyze and debias benchmarks from such
representation biases. Specifically, we generate frame-wise textual descriptions
 of videos, filter them for specific information (e.g. only objects) and leverag

e them to examine representation biases across three dimensions: 1) concept bias -- determining if a specific concept (e.g., objects) alone suffice for prediction; 2) temporal bias -- assessing if temporal information contributes to prediction; and 3) common sense vs. dataset bias -- evaluating whether zero-shot reasoning or dataset correlations contribute to prediction. We conduct a systematic analysis of 12 popular video classification and retrieval datasets and create new object-debiased test splits for these datasets. Moreover, we benchmark 30 state-of-the-art video models on original and debiased splits and analyze biases in the models. To facilitate the future development of more robust video understanding benchmarks and models, we release: "UTD-descriptions", a dataset with our rich structured descriptions for each dataset, and "UTD-splits", a dataset of object-debiased test splits.

****

Embodied Scene Understanding for Vision Language Models via MetaVQA

Weizhen Wang, Chenda Duan, Zhenghao Peng, Yuxin Liu, Bolei Zhou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22453-22464

Vision Language Models (VLMs) demonstrate significant potential as embodied AI agents for various mobility applications. However, a standardized, closed-loop benchmark for evaluating their spatial reasoning and sequential decision-making capabilities is lacking. To address this, we present MetaVQA: a comprehensive benchmark designed to assess and enhance VLMs' understanding of spatial relationships and scene dynamics through Visual Question Answering (VQA) and closed-loop simulations. MetaVQA leverages Set-of-Mark prompting and top-down view ground-truth annotations from nuScenes and Waymo datasets to automatically generate extensive question-answer pairs based on diverse real-world traffic scenarios, ensuring object-centric and context-rich instructions. Our experiments show that fine-tuning VLMs with the MetaVQA Dataset significantly improves their embodied scene understanding, which is evident not only in improved VQA accuracy but also in emerging safety-aware driving maneuvers. In addition, the learning exhibits strong transferability from simulation to real-world observation. The project webpage is at https://metadriverse.github.io/metavqa.

****

CompGS: Unleashing 2D Compositionality for Compositional Text-to-3D via Dynamically Optimizing 3D Gaussians

Chongjian Ge, Chenfeng Xu, Yuanfeng Ji, Chensheng Peng, Masayoshi Tomizuka, Ping Luo, Mingyu Ding, Varun Jampani, Wei Zhan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18509-18520

Recent breakthroughs in text-guided image generation have significantly advanced the field of 3D generation. While generating a single high-quality 3D object is now feasible, generating multiple objects with reasonable interactions within a 3D space, a.k.a. compositional 3D generation, presents substantial challenges. This paper introduces CompGS, a novel generative framework that employs 3D Gaussian Splatting (GS) for efficient, compositional text-to-3D content generation. To achieve this goal, two core designs are proposed: (1) 3D Gaussians Initialization with 2D compositionality: We transfer the well-established 2D compositionality to initialize the Gaussian parameters on an entity-by-entity basis, ensuring both consistent 3D priors for each entity and reasonable interactions among multiple entities; (2) Dynamic Optimization: We propose a dynamic strategy to optimize 3D Gaussians using Score Distillation Sampling (SDS) loss. CompGS first automatically decomposes 3D Gaussians into distinct entity parts, enabling optimization at both the entity and composition levels. Additionally, CompGS optimizes across objects of varying scales by dynamically adjusting the spatial parameters of each entity, enhancing the generation of fine-grained details, particularly in smaller entities. Qualitative comparisons and quantitative evaluations on T3Bench demonstrate the effectiveness of CompGS in generating compositional 3D objects with superior image quality and semantic alignment over existing methods. CompGS can also be easily extended to progressively adding 3D objects, facilitating complex scene generation. We hope CompGS will provide new insights to the compositional 3D generation.

****
Learning Temporally Consistent Video Depth from Video Diffusion Priors
Jiahao Shao, Yuanbo Yang, Hongyu Zhou, Youmin Zhang, Yujun Shen, Vitor Guizilini, Yue Wang, Matteo Poggi, Yiyi Liao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22841-22852
This work addresses the challenge of streamed video depth estimation, which expects not only per-frame accuracy but, more importantly, cross-frame consistency. We argue that sharing contextual information between frames or clips is pivotal in fostering temporal consistency. Therefore, we reformulate depth prediction into a conditional generation problem to provide contextual information within a clip and across clips. Specifically, we propose a consistent context-aware training and inference strategy for arbitrarily long videos to provide cross-clip context. We sample independent noise levels for each frame within a clip during training while using a sliding window strategy and initializing overlapping frames with previously predicted frames without adding noise. Moreover, we design an effective training strategy to provide context within a clip. Extensive experimental results validate our design choices and demonstrate the superiority of our approach, dubbed ChronoDepth. Project page: https://xdimlab.github.io/ChronoDepth/.
****
FIRE: Robust Detection of Diffusion-Generated Images via Frequency-Guided Reconstruction Error
Beilin Chu, Xuan Xu, Xin Wang, Yufei Zhang, Weike You, Linna Zhou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12830-12839
The rapid advancement of diffusion models has significantly improved high-quality image generation, making generated content increasingly challenging to distinguish from real images and raising concerns about potential misuse. In this paper, we observe that diffusion models struggle to accurately reconstruct mid-band frequency information in real images, suggesting the limitation could serve as a cue for detecting diffusion model generated images. Motivated by this observation, we propose a novel method called Frequency-guIded Reconstruction Error (FIRE), which, to the best of our knowledge, is the first to investigate the influence of frequency decomposition on reconstruction error. FIRE assesses the variation in reconstruction error before and after the frequency decomposition, offering a robust method for identifying diffusion model generated images. Extensive experiments show that FIRE generalizes effectively to unseen diffusion models and maintains robustness against diverse perturbations.
****
Assessing and Learning Alignment of Unimodal Vision and Language Models
Le Zhang, Qian Yang, Aishwarya Agrawal; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14604-14614
How well are unimodal vision and language models aligned? While prior work has explored this question, their assessment methods do not directly translate to practical vision-language tasks. In this paper, we propose a direct assessment method, inspired by linear probing, to evaluate vision-language alignment. We identify that the degree of alignment of SSL vision models depends on their SSL training objective and find that clustering quality of SSL representations impacts alignment performance more than their linear separability. We then introduce Swift Alignment of Image and Language (SAIL), an efficient transfer learning framework that aligns pretrained unimodal vision and language models for downstream tasks. SAIL requires significantly less paired image-text data (6%) compared to models like CLIP, which are trained from scratch. It trains with a single A100 GPU in 5 hours and supports a batch size of up to 32,768. SAIL achieves 73.4% zero-shot accuracy on ImageNet (compared to CLIP's 72.7%) and excels in zero-shot retrieval, complex reasoning, and semantic segmentation. SAIL also enhances the language-compatibility of vision encoders, improving the performance of multimodal large language models. The full codebase and model weights are open-source: https://lezhang7.github.io/sail.github.io/
****
Samba: A Unified Mamba-based Framework for General Salient Object Detection

Jiahao He, Keren Fu, Xiaohong Liu, Qijun Zhao; Proceedings of the Computer Visio n and Pattern Recognition Conference (CVPR), 2025, pp. 25314-25324
Existing salient object detection (SOD) models primarily resort to convolutional neural networks (CNNs) and Transformers. However, the limited receptive fields of CNNs and quadratic computational complexity of transformers both constrain th e performance of current models on discovering attention-grabbing objects. The e merging state space model, namely Mamba, has demonstrated its potential to balan ce global receptive fields and computational complexity. Therefore, we propose a novel unified framework based on the pure Mamba architecture, dubbed saliency M amba (Samba), to flexibly handle general SOD tasks, including RGB/RGB-D/RGB-T SO D, video SOD (VSOD), and RGB-D VSOD. Specifically, we rethink Mamba's scanning s trategy from the perspective of SOD, and identify the importance of maintaining spatial continuity of salient patches within scanning sequences. Based on this, we propose a saliency-guided Mamba block (SGMB), incorporating a spatial neighbo ring scanning (SNS) algorithm to preserve spatial continuity of salient patches. Additionally, we propose a context-aware upsampling (CAU) method to promote hie rarchical feature alignment and aggregations by modeling contextual dependencies . Experimental results show that our Samba outperforms existing methods across f ive SOD tasks on 21 datasets with lower computational cost, confirming the super iority of introducing Mamba to the SOD areas. Our code is available at https:// github.com/Jia-hao999/Samba.
****

Action Detail Matters: Refining Video Recognition with Local Action Queries
Mengmeng Wang, Zeyi Huang, Xiangjie Kong, Guojiang Shen, Guang Dai, Jingdong Wan g, Yong Liu; Proceedings of the Computer Vision and Pattern Recognition Conferen ce (CVPR), 2025, pp. 19132-19142
Video action recognition involves interpreting both global context and specific details to accurately identify actions. While previous models are effective at c apturing spatiotemporal features, they often lack a focused representation of ke y action details. To address this, we introduce \nameo, a framework designed for refining video action recognition through integrated global and local feature l earning. Inspired by human visual cognition theory, our approach balances the fo cus on both broad contextual changes and action-specific details, minimizing the influence of irrelevant background noise. We first employ learnable action quer ies to selectively emphasize action-relevant regions without requiring region-sp ecific labels. Next, these queries are learned by a local action streaming branc h that enables progressive query propagation. Moreover, we introduce a parameter -free feature interaction mechanism for effective multi-scale interaction betwee n global and local features with minimal additional overhead. Extensive experime nts demonstrate that \name achieves state-of-the-art performance across multiple action recognition datasets, validating its effectiveness and robustness in han dling action-relevant details.
****

PAVE: Patching and Adapting Video Large Language Models
Zhuoming Liu, Yiquan Li, Khoi Duc Nguyen, Yiwu Zhong, Yin Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3306-3317
We present PAVE, a framework for adapting pre-trained video large language model s (Video-LLMs) to downstream tasks that incorporate side-channel signals, such a s audio, camera pose, or high frame rate videos. PAVE introduces a lightweight a daptation strategy called "patching", which adds a small number of parameters an d operations to the base model without modifying its architecture or pre-trained weights. We demonstrate that PAVE effectively enhances pre-trained Video-LLMs w ith the cost of adding <1% additional FLOPs and parameters for diverse tasks, in cluding audio-visual understanding, 3D reasoning, and multi-view video understan ding, surpassing state-of-the-art task-specific models. Moreover, when applied t o high frame rate videos, PAVE further improves video understanding, enhancing t he performance of strong base models. Finally, our experiments show that our fra mework generalizes well across different Video-LLMs.
****

LesionLocator: Zero-Shot Universal Tumor Segmentation and Tracking in 3D Whole-B

ody Imaging
Maximilian Rokuss, Yannick Kirchhoff, Seval Akbal, Balint Kovacs, Saikat Roy, Co nstantin Ulrich, Tassilo Wald, Lukas T. Rotkopf, Heinz-Peter Schlemmer, Klaus Ma ier-Hein; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30872-30885
In this work, we present LesionLocator, a framework for zero-shot longitudinal l esion tracking and segmentation in 3D medical imaging, establishing the first en d-to-end model capable of 4D tracking with dense spatial prompts. Our model leve rages an extensive dataset of 23,262 annotated medical scans, as well as synthes ized longitudinal data across diverse lesion types. The diversity and scale of o ur dataset significantly enhances model generalizability to real-world medical i maging challenges and addresses key limitations in longitudinal data availabilit y. LesionLocator outperforms all existing promptable models in lesion segmentati on by nearly 10 dice points, reaching human-level performance, and achieves stat e-of-the-art results in lesion tracking, with superior lesion retrieval and segm entation accuracy. LesionLocator not only sets a new benchmark in universal prom ptable lesion segmentation and automated longitudinal lesion tracking but also p rovides the first open-access solution of its kind, releasing our synthetic 4D d ataset and model to the community, empowering future advancements in medical ima ging. Code is available at: www.github.com/MIC-DKFZ/LesionLocator
****

Generative Map Priors for Collaborative BEV Semantic Segmentation
Jiahui Fu, Yue Gong, Luting Wang, Shifeng Zhang, Xu Zhou, Si Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11919-11928
Collaborative perception aims to address the constraint of single-agent percepti on by exchanging information among multiple agents. Previous works primarily foc us on collaborative object detection, exploring compressed transmission and fusi on prediction tailored to sparse object features. However, these strategies are not well-suited for dense features in collaborative BEV semantic segmentation. T herefore, we propose CoGMP, a novel Collaborative framework that leverages Gener ative Map Priors to achieve efficient compression and robust fusion. CoGMP intro duces two key innovations: Element Format Feature Compression (EFFC) and Structu re Guided Feature Fusion (SGFF). Specifically, EFFC leverages map element priors from codebook to encode BEV features as discrete element indices for transmitte d information compression. Meanwhile, SGFF utilizes a diffusion model with struc tural priors to coherently integrate multi-agent features, thereby achieving con sistent fusion predictions. Evaluations on the OPV2V dataset show that CoGMP ach ieves a 6.89/7.64 Road/Lane IoU improvement and a 32-fold reduction in communica tion volume.
****

Coherent 3D Portrait Video Reconstruction via Triplane Fusion
Shengze Wang, Xueting Li, Chao Liu, Matthew Chan, Michael Stengel, Henry Fuchs, Shalini De Mello, Koki Nagano; Proceedings of the Computer Vision and Pattern Re cognition Conference (CVPR), 2025, pp. 10712-10722
Recent breakthroughs in single-image 3D portrait reconstruction have enabled tel epresence systems to stream 3D portrait videos from a single camera in real-time , democratizing telepresence. However, per-frame 3D reconstruction exhibits temp oral inconsistency and forgets the user's appearance. On the other hand, self-re enactment methods can render coherent 3D portraits by driving a 3D avatar built from a single reference image but fail to faithfully preserve the user's per-fra me appearance (e.g., instantaneous facial expressions and lighting). As a result , neither of these two frameworks is an ideal solution for democratized 3D telep resence. In this work, we address this dilemma and propose a novel solution that maintains both coherent identity and dynamic per-frame appearance to enable the best possible realism. To this end, we propose a new fusion-based method that t akes the best of both worlds by fusing a canonical 3D prior from a reference vie w with dynamic appearance from per-frame input views, producing temporally stabl e 3D videos with faithful reconstruction of the user's per-frame appearance. Tra ined only using synthetic data produced by an expression-conditioned 3D GAN, our

encoder-based method achieves both state-of-the-art 3D reconstruction and temporal consistency on in-studio and in-the-wild datasets.
****

Generative Image Layer Decomposition with Visual Effects
Jinrui Yang, Qing Liu, Yijun Li, Soo Ye Kim, Daniil Pakhomov, Mengwei Ren, Jianming Zhang, Zhe Lin, Cihang Xie, Yuyin Zhou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7643-7653
Recent advancements in large generative models, particularly diffusion-based methods, have significantly enhanced the capabilities of image editing. However, achieving precise control over image composition tasks remains a challenge. Layered representations, which allow for independent editing of image components, are essential for user-driven content creation, yet existing approaches often struggle to decompose image into plausible layers with accurately retained transparent visual effects such as shadows and reflections. We propose LayerDecomp, a generative framework for image layer decomposition which outputs photorealistic clean backgrounds and high-quality transparent foregrounds with faithfully preserved visual effects. To enable effective training, we first introduce a dataset preparation pipeline that automatically scales up simulated multi-layer data with synthesized visual effects. To further enhance real-world applicability, we supplement this simulated dataset with camera-captured images containing natural visual effects. Additionally, we propose a consistency loss which enforces the model to learn accurate representations for the transparent foreground layer when ground-truth annotations are not available. Our method achieves superior quality in layer decomposition, outperforming existing approaches in object removal and spatial editing tasks across several benchmarks and multiple user studies, unlocking various creative possibilities for layer-wise image editing.
****

AR-Diffusion: Asynchronous Video Generation with Auto-Regressive Diffusion
Mingzhen Sun, Weining Wang, Gen Li, Jiawei Liu, Jiahui Sun, Wanquan Feng, Shanshan Lao, Siyu Zhou, Qian He, Jing Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7364-7373
The task of video generation requires synthesizing visually realistic and temporally coherent video frames. Existing methods primarily use asynchronous auto-regressive models or synchronous diffusion models to address this challenge. However, asynchronous auto-regressive models often suffer from inconsistencies between training and inference, leading to issues such as error accumulation, while synchronous diffusion models are limited by their reliance on rigid sequence length. To address these issues, we introduce Auto-Regressive Diffusion (AR-Diffusion), a novel model that combines the strengths of auto-regressive and diffusion models for flexible, asynchronous video generation. Specifically, our approach leverages diffusion to gradually corrupt video frames in both training and inference, reducing the discrepancy between these phases. Inspired by auto-regressive generation, we incorporate a non-decreasing constraint on the corruption timesteps of individual frames, ensuring that earlier frames remain clearer than subsequent ones. This setup, together with temporal causal attention, enables flexible generation of videos with varying lengths while preserving temporal coherence. In addition, we design two specialized timestep schedulers: the FoPP scheduler for balanced timestep sampling during training, and the AD scheduler for flexible timestep differences during inference, supporting both synchronous and asynchronous generation. Extensive experiments demonstrate the superiority of our proposed method, by achieving competitive and state-of-the-art results across four challenging benchmarks.
****

ManiVideo: Generating Hand-Object Manipulation Video with Dexterous and Generalizable Grasping
Youxin Pang, Ruizhi Shao, Jiajun Zhang, Hanzhang Tu, Yun Liu, Boyao Zhou, Hongwen Zhang, Yebin Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12209-12219
In this paper, we introduce ManiVideo, a novel method for generating consistent and temporally coherent bimanual hand-object manipulation videos from given moti

on sequences of hands and objects. The core idea of ManiVideo is the constructio n of a multi-layer occlusion (MLO) representation that learns 3D occlusion relat ionships from occlusion-free normal maps and occlusion confidence maps. By embed ding the MLO structure into the UNet in two forms, the model enhances the 3D con sistency of dexterous hand-object manipulation. To further achieve the generaliz able grasping of objects, we integrate Objaverse, a large-scale 3D object datase t, to address the scarcity of video data, thereby facilitating the learning of e xtensive object consistency. Additionally, we propose an innovative training str ategy that effectively integrates multiple datasets, supporting downstream tasks such as human-centric hand-object manipulation video generation. Through extens ive experiments, we demonstrate that our approach not only achieves video genera tion with plausible hand-object interaction and generalizable objects, but also outperforms existing SOTA methods.

****

DOF-GS: Adjustable Depth-of-Field 3D Gaussian Splatting for Post-Capture Refocus ing, Defocus Rendering and Blur Removal

Yujie Wang, Praneeth Chakravarthula, Baoquan Chen; Proceedings of the Computer V ision and Pattern Recognition Conference (CVPR), 2025, pp. 21297-21306

3D Gaussian Splatting (3DGS) techniques have recently enabled high-quality 3D sc ene reconstruction and real-time novel view synthesis. These approaches, however , are limited by the pinhole camera model and lack effective modeling of defocus effects. Departing from this, we introduce DOF-- a new 3DGS-based framework wit h a finite-aperture camera model and explicit, differentiable defocus rendering, enabling it to function as a post-capture control tool. By training with multi- view images with moderate defocus blur, DOF-GS learns inherent camera characteri stics and reconstructs sharp details of the underlying scene, particularly, enab ling rendering of varying DOF effects through on-demand aperture and focal dista nce control, post-capture and optimization. Additionally, our framework extracts circle-of-confusion cues during optimization to identify in-focus regions in in put views, enhancing the reconstructed 3D scene details. Experimental results de monstrate that DOF-GS supports post-capture refocusing, adjustable defocus and h igh-quality all-in-focus rendering, from multi-view images with uncalibrated def ocus blur.

****

The Photographer's Eye: Teaching Multimodal Large Language Models to See, and Cr itique Like Photographers

Daiqing Qi, Handong Zhao, Jing Shi, Simon Jenni, Yifei Fan, Franck Dernoncourt, Scott Cohen, Sheng Li; Proceedings of the Computer Vision and Pattern Recognitio n Conference (CVPR), 2025, pp. 24807-24816

Photographer, curator, and former director of photography at the Museum of Moder n Art (MoMA), John Szarkowski remarked in *William Eggleston's Guide*, "While ed iting directly from life, photographers have found it too difficult to see simul taneously both the blue and the sky." Szarkowski insightfully revealed a notable gap between general and aesthetic visual understanding: while the former emphas izes identifying factual elements in an image (the sky), the latter transcends m ere object identification, viewing it instead as an aesthetic component--a pure expanse of blue, valued purely as a color block in visual aesthetics. Such disti nctions between general visual understanding (detection, localization, etc.) and aesthetic perception (color, lighting, composition, etc.) pose a significant ch allenge for existing Multimodal Large Language Models (MLLMs) in comprehending i mage aesthetics, which is increasingly needed in real-world applications, from i mage recommendation and enhancement to generation. To fundamentally advance the aesthetic understanding of MLLMs, we introduce a novel dataset, PhotoCritique, d erived from extensive discussions among professional photographers and enthusias ts, distinguished by its large scale, expertise, and diversity. Additionally, we propose a new model, PhotoEye, an MLLM featuring a language-guided multi-view v ision fusion mechanism for understanding image aesthetics from multiple perspect ives. Finally, we introduce PhotoBench, a comprehensive and professional benchma rk for aesthetic visual understanding. Our model demonstrates significant advant ages over both open-source and commercial models on existing benchmarks and Phot

oBench.
****
Revisiting Audio-Visual Segmentation with Vision-Centric Transformer
Shaofei Huang, Rui Ling, Tianrui Hui, Hongyu Li, Xu Zhou, Shifeng Zhang, Si Liu, Richang Hong, Meng Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8352-8361
Audio-Visual Segmentation (AVS) aims to segment sound-producing objects in video frames based on the associated audio signal. Prevailing AVS methods typically adopt an audio-centric Transformer architecture, where object queries are derived from audio features. However, audio-centric Transformers suffer from two limitations: perception ambiguity caused by the mixed nature of audio, and weakened dense prediction ability due to visual detail loss. To address these limitations, we propose a new Vision-Centric Transformer (VCT) framework that leverages vision-derived queries to iteratively fetch corresponding audio and visual information, enabling queries to better distinguish between different sounding objects from mixed audio and accurately delineate their contours. Additionally, we also introduce a Prototype Prompted Query Generation (PPQG) module within our VCT framework to generate vision-derived queries that are both semantically aware and visually rich through audio prototype prompting and pixel context grouping, facilitating audio-visual information aggregation. Extensive experiments demonstrate that our VCT framework achieves new state-of-the-art performances on three subsets of the AVSBench dataset.
****
Synergizing Motion and Appearance: Multi-Scale Compensatory Codebooks for Talking Head Video Generation
Shuling Zhao, Fa-Ting Hong, Xiaoshui Huang, Dan Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26232-26241
Talking head video generation aims to generate a realistic talking head video that preserves the person's identity from a source image and the motion from a driving video. Despite the promising progress made in the field, it remains a challenging and critical problem to generate videos with accurate poses and fine-grained facial details simultaneously. Essentially, facial motion is often highly complex to model precisely, and the one-shot source face image cannot provide sufficient appearance guidance during generation due to dynamic pose changes. To tackle the problem, we propose to jointly learn motion and appearance codebooks and perform multi-scale codebook compensation to effectively refine both the facial motion conditions and appearance features for talking face image decoding. Specifically, the designed multi-scale motion and appearance codebooks are learned simultaneously in a unified framework to store representative global facial motion flow and appearance patterns. Then, we present a novel multi-scale motion and appearance compensation module, which utilizes a transformer-based codebook retrieval strategy to query complementary information from the two codebooks for joint motion and appearance compensation. The entire process produces motion flows of greater flexibility and appearance features with fewer distortions across different scales, resulting in a high-quality talking head video generation framework. Extensive experiments on various benchmarks validate the effectiveness of our approach and demonstrate superior generation results from both qualitative and quantitative perspectives when compared to state-of-the-art competitors. The project page is available at https://shaelynz.github.io/synergize-motion-appearance/.
****
HOIGPT: Learning Long-Sequence Hand-Object Interaction with Language Models
Mingzhen Huang, Fu-Jen Chu, Bugra Tekin, Kevin J. Liang, Haoyu Ma, Weiyao Wang, Xingyu Chen, Pierre Gleize, Hongfei Xue, Siwei Lyu, Kris Kitani, Matt Feiszli, Hao Tang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7136-7146
We introduce HOIGPT, a token-based generative method that unifies 3D hand-object interactions (HOI) perception and generation, offering the first comprehensive solution for captioning and generating high-quality 3D HOI sequences from a diverse range of conditional signals (e.g. text, objects, partial sequences). At its

core, HOIGPT utilizes a large language model to predict the bidirectional transformation between HOI sequences and natural language descriptions. Given text inputs, HOIGPT generates a sequence of hand and object meshes; given (partial) HOI sequences, HOIGPT generates text descriptions and completes the sequences. To facilitate HOI understanding with a large language model, this paper introduces two key innovations: (1) a novel physically grounded HOI tokenizer, the hand-object decomposed VQ-VAE, for discretizing HOI sequences, and (2) a motion-aware language model trained to process and generate both text and HOI tokens. Extensive experiments demonstrate that HOIGPT sets new state-of-the-art performance on both text generation (+2.01% R Precision) and HOI generation (-2.56 FID) across multiple tasks and benchmarks.

****

GraphI2P: Image-to-Point Cloud Registration with Exploring Pattern of Correspondence via Graph Learning

Lin Bie, Shouan Pan, Siqi Li, Yining Zhao, Yue Gao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22161-22171

Although the fusion of images and LiDAR point clouds is crucial to many applications in computer vision, the relative poses of cameras and LiDAR scanners are often unknown. The general registration pipeline first establishes correspondences and then performs pose estimation based on the generated matches. However, 2D-3D correspondences are inherently challenging to establish due to the large gap between images and LiDAR point clouds. To this end, we build a bridge to alleviate the 2D-3D gap and propose a practical framework to align LiDAR point clouds to the virtual points generated by images. In this way, the modality gap is converted to the domain gap of point clouds. Moreover, we propose a virtual-spherical representation and adaptive distribution sample module to narrow the domain gap between virtual and LiDAR point clouds. Then, we explore the reliable correspondence pattern consistency through a graph-based selection process. We improve the correspondence representation through a graph neural network. Experimental results demonstrate that our method outperforms the state-of-the-art methods by more than 10.77% and 12.53% performance on the KITTI Odometry and nuScenes datasets, respectively. The results demonstrate that our method can effectively solve non-synchronized random frame registration.

****

SoftVQ-VAE: Efficient 1-Dimensional Continuous Tokenizer

Hao Chen, Ze Wang, Xiang Li, Ximeng Sun, Fangyi Chen, Jiang Liu, Jindong Wang, Bhiksha Raj, Zicheng Liu, Emad Barsoum; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28358-28370

Efficient image tokenization with high compression ratios remains a critical challenge for training generative models.We present SoftVQ-VAE, a continuous image tokenizer that leverages soft categorical posteriors to aggregate multiple codewords into each latent token, substantially increasing the representation capacity of the latent space. When applied to Transformer-based architectures, our approach compresses 256x256 and 512x512 images using only 32 or 64 1-dimensional tokens.Not only does SoftVQ-VAE show consistent and high-quality reconstruction, more importantly, it also achieves state-of-the-art and significantly faster image generation results across different denoising-based generative models. Remarkably, SoftVQ-VAE improves inference throughput by up to 18x for generating 256x256 images and 55x for 512x512 images while achieving competitive FID scores of 1.78 and 2.21 for SiT-XL.It also improves the training efficiency of the generative models by reducing the number of training iterations by 2.3x while maintaining comparable performance. With its fully-differentiable design and semantic-rich latent space, our experiment demonstrates that SoftVQ-VQE achieves efficient tokenization without compromising generation quality, paving the way for more efficient generative models.Code and model will be released.

****

FedCS: Coreset Selection for Federated Learning

Chenhe Hao, Weiying Xie, Daixun Li, Haonan Qin, Hangyu Ye, Leyuan Fang, Yunsong Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15434-15443

Federated Learning (FL) is an emerging direction in distributed machine learning that enables jointly training a model without sharing the data. However, as the size of datasets grows exponentially, computational costs of FL increase. In this paper, we propose the first Coreset Selection criterion for Federated Learning (FedCS) by exploring the Distance Contrast (DC) in feature space. Our FedCS is inspired by the discovery that DC can indicate the intrinsic properties inherent to samples regardless of the networks. Based on the observation, we develop a method that is mathematically formulated to prune samples with high DC. The principle behind our pruning is that high DC samples either contain less information or represent rare extreme cases, thus removal of them can enhance the aggregation performance. Besides, we experimentally show that samples with low DC usually contain substantial information and reflect the common features of samples within their classes, such that they are suitable for constructing coreset. With only two time of linear-logarithmic complexity operation, FedCS leads to significant improvements over the methods using whole dataset in terms of computational costs, with similar accuracies. For example, on the CIFAR-10 dataset with Dirichlet coefficient \alpha=0.1, FedCS achieves 58.88% accuracy using only 44% of the entire dataset, whereas other methods require twice the data volume as FedCS for same performance.

****

DPC: Dual-Prompt Collaboration for Tuning Vision-Language Models

Haoyang Li, Liang Wang, Chao Wang, Jing Jiang, Yan Peng, Guodong Long; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25623-25632

The Base-New Trade-off (BNT) problem universally exists during the optimization of CLIP-based prompt tuning, where continuous fine-tuning on base (target) classes leads to a simultaneous decrease of generalization ability on new (unseen) classes. Existing approaches attempt to regulate the prompt tuning process to balance BNT by appending constraints. However, imposed on the same target prompt, these constraints fail to fully avert the mutual exclusivity between the optimization directions for base and new. As a novel solution to this challenge, we propose the plug-and-play Dual-Prompt Collaboration (DPC) framework, the first that decoupling the optimization processes of base and new tasks at the prompt level. Specifically, we clone a learnable parallel prompt based on the backbone prompt, and introduce a variable Weighting-Decoupling framework to independently control the optimization directions of dual prompts specific to base or new tasks, thus avoiding the conflict in generalization. Meanwhile, we propose a Dynamic Hard Negative Optimizer, utilizing dual prompts to construct a more challenging optimization task on base classes for enhancement. For interpretability, we prove the feature channel invariance of the prompt vector during the optimization process, providing theoretical support for the Weighting-Decoupling of DPC. Extensive experiments on multiple backbones demonstrate that DPC can significantly improve base performance without introducing any external knowledge beyond the base classes, while maintaining generalization to new classes. Code is available at: https://github.com/JREion/DPC.

****

Dual-Granularity Semantic Guided Sparse Routing Diffusion Model for General Pansharpening

Yinghui Xing, Litao Qu, Shizhou Zhang, Di Xu, Yingkun Yang, Yanning Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12658-12668

Pansharpening aims at integrating complementary information from panchromatic and multispectral images. Available deep-learning based pansharpening methods typically perform exceptionally with particular satellite datasets. At the same time, it has been observed that these models also exhibit scene dependence, for example, if the majority of the training samples come from the urban scenes, the model's performance may decline in the river scene. To address the domain gap produced by varying satellite sensors and distinct scenes, we propose a dual-granularity semantic guided sparse routing diffusion model for general pansharpening. By utilizing the large Vision-Language Models (VLMs) in the field of geoscience, e

.g, GeoChat, we introduce the dual granularity semantics to generate dynamic spa rse routing scores for adaptation of different satellite sensors and scenes. Thi s scene-level and region-level dual-granularity semantic information serves as g uidance for dynamically activating specialized experts within the diffusion mode l. Extensive experiments on WorldView-3, QuickBird, and GaoFen-2 datasets show t he effectiveness of our proposed method. Notably, the proposed method outperform s the comparison approaches in adapting to new satellite sensors and scenes. The codes are available at https://github.com/codgodtao/SGDiff.

****

AIM-Fair: Advancing Algorithmic Fairness via Selectively Fine-Tuning Biased Mode ls with Contextual Synthetic Data

Zengqun Zhao, Ziquan Liu, Yu Cao, Shaogang Gong, Ioannis Patras; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28748-2 8758

Recent advances in generative models have sparked research on improving model fa irness with AI-generated data. However, existing methods often face limitations in the diversity and quality of synthetic data, leading to compromised fairness and overall model accuracy. Moreover, many approaches rely on the availability o f demographic group labels, which are often costly to annotate. This paper propo ses AIM-Fair, aiming to overcome these limitations and harness the potential of cutting-edge generative models in promoting algorithmic fairness. We investigate a fine-tuning paradigm starting from a biased model initially trained on real-w orld data without demographic annotations. This model is then fine-tuned using u nbiased synthetic data generated by a state-of-the-art diffusion model to improv e its fairness. Two key challenges are identified in this fine-tuning paradigm, 1) the low quality of synthetic data, which can still happen even with advanced generative models, and 2) the domain and bias gap between real and synthetic dat a. To address the limitation of synthetic data quality, we propose Contextual Sy nthetic Data Generation (CSDG) to generate data using a text-to-image diffusion model (T2I) with prompts generated by a context-aware LLM, ensuring both data di versity and control of bias in synthetic data. To resolve domain and bias shifts , we introduce a novel selective fine-tuning scheme in which only model parameter s more sensitive to bias and less sensitive to domain shift are updated. Experi ments on CelebA and UTKFace datasets show that our AIM-Fair improves model fairn ess while maintaining utility, outperforming both fully and partially fine-tuned approaches to model fairness.

****

Robust Multi-Object 4D Generation for In-the-wild Videos

Wen-Hsuan Chu, Lei Ke, Jianmeng Liu, Mingxiao Huo, Pavel Tokmakov, Katerina Frag kiadaki; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22067-22077

We address the challenge of generating dynamic 4D scenes from monocular multi-ob ject videos with heavy occlusions and introduce Robust4DGen, a novel approach th at integrates rendering-based deformable 3D Gaussian optimization with generativ e priors for view synthesis. While existing view-synthesis models excel at novel view generation for isolated objects, they struggle with full scenes due to the ir complexity and data demands. To overcome this, Robust4DGen decomposes scenes into individual objects, optimizing a differentiable set of deformable Gaussians per object while capturing 2D occlusions from a 3D perspective through joint Ga ussian splatting. Joint splatting ensures occlusion-aware rendering losses in ob served frames while explicit object decomposition allows the usage of object-cen tric diffusion models for object completion in unobserved viewpoints. To reconci le the differences between object-centric priors and the global frame-centric co ordinate system of the video, Robust4DGen employs differentiable transformations to unify the rendering and generative constraints within a single framework. Th e result is a model capable of generating 4D objects across space and time while producing 2D and 3D point tracks from monocular videos. To rigorously evaluate the quality of scene generation and the accuracy of the motion under multi-objec t occlusions, we introduce MOSE-PTS, a subset of the challenging MOSE benchmark, which we annotated with high-quality 2D point tracks. Quantitative evaluations

and perceptual human studies confirm that Robust4DGen generates more realistic n
ovel views of scenes and produces more accurate point tracks compared to existin
g approaches.
****

OmniMMI: A Comprehensive Multi-modal Interaction Benchmark in Streaming Video Co
ntexts

Yuxuan Wang, Yueqian Wang, Bo Chen, Tong Wu, Dongyan Zhao, Zilong Zheng; Proceed
ings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp.
 18925-18935

The rapid advancement of multi-modal language models (MLLMs) like GPT-4o has pro
pelled the development of Omni language models, designed to process and proactiv
ely respond to continuous streams of multi-modal data. Despite their potential,
evaluating their real-world interactive capabilities in streaming video contexts
 remains a formidable challenge. In this work, we introduce OmniMMI, a comprehen
sive multi-modal interaction benchmark tailored for OmniLLMs in streaming video
contexts. OmniMMI encompasses over 1,121 videos and 2,290 questions, addressing
two critical yet underexplored challenges in existing video benchmarks: streamin
g video understanding and proactive reasoning, across six distinct subtasks. Mo
reover, we propose a novel framework, Multi-modal Multiplexing Modeling (M4), de
signed to enable an inference-efficient streaming model that can see, listen whi
le generating.
****

SOAP: Vision-Centric 3D Semantic Scene Completion with Scene-Adaptive Decoder an
d Occluded Region-Aware View Projection

Hyo-Jun Lee, Yeong Jun Koh, Hanul Kim, Hyunseop Kim, Yonguk Lee, Jinu Lee; Proce
edings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, p
p. 17145-17154

Existing view transformations in vision-centric 3D Semantic Scene Completion (SS
C) inevitably experience erroneous feature duplication in the reconstructed voxe
l space due to occlusions, leading to a dilution of informative contexts. Furthe
rmore, semantic classes exhibit high variability in their appearance in real-wor
ld driving scenarios. To address these issues, we introduce a novel 3D SSC metho
d, called SOAP, including two key components: an occluded region-aware view proj
ection and a scene-adaptive decoder. The occluded region-aware view projection e
ffectively converts 2D image features into voxel space, refining the duplicated
features of occluded regions using information gathered from previous observatio
ns. The scene-adaptive decoder guides query embeddings to learn diverse driving
environments based on a comprehensive semantic repository. Extensive experiments
 validate that the proposed SOAP significantly outperforms existing methods for
the vision-centric 3D SSC on automated driving datasets, SemanticKITTI and SSCBe
nch. Code is available at https://github.com/gywns6287/SOAP.
****

Taxonomy-Aware Evaluation of Vision-Language Models

Vésteinn Snæbjarnarson, Kevin Du, Niklas Stoehr, Serge Belongie, Ryan Cotterell,
 Nico Lang, Stella Frank; Proceedings of the Computer Vision and Pattern Recogni
tion Conference (CVPR), 2025, pp. 9109-9120

When a vision-language model (VLM) is prompted to identify an entity depicted in
 an image, it may answer "I see a conifer," rather than the specific label "Norw
ay spruce". This raises two issues for evaluation: Firstly, the unconstrained ge
nerated text needs to be mapped to the evaluation label space (i.e., "conifer").
 Secondly, a useful classification measure should give partial credit to less sp
ecific, but not incorrect, answers ("Norway spruce" being a type of "conifer").
To meet these requirements, we propose a framework for evaluating unconstrained
text predictions such as those generated from a vision-language model against a
taxonomy. Specifically, we propose the use of hierarchical precision and recall
measures to assess the level of correctness and specificity of predictions with
regard to a taxonomy. Experimentally, we first show that existing text similarit
y measures do not capture taxonomic similarity well. We then develop and compare
 different methods to map textual VLM predictions onto a taxonomy. This allows u
s to compute hierarchical similarity measures between the generated text and the

ground truth labels. Finally, we analyze modern VLMs on fine-grained visual classification tasks based on our proposed taxonomic evaluation scheme.
****

Active Event-based Stereo Vision
Jianing Li, Yunjian Zhang, Haiqian Han, Xiangyang Ji; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 971-981
Conventional frame-based imaging for active stereo systems has encountered major challenges in fast-motion scenarios. However, how to design a novel paradigm for high-speed depth sensing still remains an open issue. In this paper, we propose a novel problem setting, namely active event-based stereo vision, which provides the first insight of integrating binocular event cameras and an infrared projector for high-speed depth sensing. Technically, we first build a stereo camera prototype system and present a real-world dataset with over 21.5k spatiotemporal synchronized labels at 15 Hz, while also creating a realistic synthetic dataset with stereo event streams and 23.8k synchronized labels at 20 Hz. Then, we propose ActiveEventNet, a lightweight yet effective active event-based stereo matching neural network that learns to generate high-quality dense disparity maps from stereo event streams with low latency. Experiments demonstrate that our ActiveEventNet outperforms state-of-the-art methods meanwhile significantly reducing computational complexity. Our solution offers superior depth sensing compared to conventional stereo cameras in high-speed scenes, while also achieving the inference speed of up to 150 FPS with our prototype. We believe that this novel paradigm will provide new insights into future depth sensing systems. Our project can be available at https://github.com/jianing-li/active_event_based_stereo.
****

Mono-InternVL: Pushing the Boundaries of Monolithic Multimodal Large Language Models with Endogenous Visual Pre-training
Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jiawen Liu, Jifeng Dai, Yu Qiao, Xizhou Zhu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24960-24971
In this paper, we focus on monolithic Multimodal Large Language Models (MLLMs) that integrate visual encoding and language decoding into a single LLM. In particular, we identify that existing pre-training strategies for monolithic MLLMs often suffer from unstable optimization or catastrophic forgetting. To address this issue, our core idea is to embed a new visual parameter space into a pre-trained LLM, thereby stably learning visual knowledge from noisy data while freezing the LLM. Based on this principle, we present Mono-InternVL, a novel monolithic MLLM that seamlessly integrates a set of visual experts via a multimodal mixture-of-experts structure. Moreover, we propose an innovative pre-training strategy to maximize the visual capability of Mono-InternVL, namely Endogenous Visual Pre-training (EViP). In particular, EViP is designed as a progressive learning process for visual experts, which aims to fully exploit the visual knowledge from noisy data to high-quality data. To validate our approach, we conduct extensive experiments on 16 benchmarks. Experimental results confirm the superior performance of Mono-InternVL over existing monolithic MLLMs on 13 of 16 multimodal benchmarks, e.g., +80 points over Emu3 on OCRBench. Compared to the modular baseline, i.e., InternVL-1.5, Mono-InternVL still retains comparable multimodal performance while reducing up to 67% first token latency. Our project is available at https://internvl.github.io/blog/2024-10-10-Mono-InternVL/.
****

SimVS: Simulating World Inconsistencies for Robust View Synthesis
Alex Trevithick, Roni Paiss, Philipp Henzler, Dor Verbin, Rundi Wu, Hadi Alzayer, Ruiqi Gao, Ben Poole, Jonathan T. Barron, Aleksander Holynski, Ravi Ramamoorthi, Pratul P. Srinivasan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16464-16474
Novel-view synthesis techniques achieve impressive results for static scenes but struggle when faced with the inconsistencies inherent to casual capture settings: varying illumination, scene motion, and other unintended effects that are difficult to model explicitly. We present an approach for leveraging generative video models to simulate the inconsistencies in the world that can occur during cap

ture. We use this process, along with existing multi-view datasets, to create synthetic data for training a multi-view harmonization network that is able to reconcile inconsistent observations into a consistent 3D scene. We demonstrate that our world-simulation strategy significantly outperforms traditional augmentation methods in handling real-world scene variations, thereby enabling highly accurate static 3D reconstructions in the presence of a variety of challenging inconsistencies.
****

FLAVC: Learned Video Compression with Feature Level Attention
Chun Zhang, Heming Sun, Jiro Katto; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28019-28028
Learned Video Compression (LVC) aims to reduce redundancy in sequential data through deep learning approaches. Recent advances have significantly boosted LVC performance by shifting compression operations to feature domain, often combining Motion Estimation and Motion Compensation module(MEMC) with CNN-based context extraction. However, reliance on motions and convolution-driven context models limits its generalizability and global perception. To address these issues, we propose a Feature-level Attention (FLA) module within a Transformer-based framework that perceives full-frame explicitly, thus bypassing confined motion signatures. FLA accomplishes global perception by converting high-level local patch embeddings to one-dimensional batch-wise vectors and replacing traditional attention weights to a global context matrix. Amongst this, a dense overlapping patcher (DP) is introduced to retain local features before embedding projection. Furthermore, a Transformer-CNN mixed encoder is applied to alleviate the spatial feature bottleneck without expanding latent size. Experiments demonstrate excellent generalizability with universally efficient redundancy reduction in different scenarios. Extensive tests on four video compression datasets show that our method achieves state-of-the-art Rate-Distortion performance compared to existing LVC methods and traditional codecs. A down-scaled version of our model reduced computation overhead by a great margin while maintained great performance.
****

An End-to-End Robust Point Cloud Semantic Segmentation Network with Single-Step Conditional Diffusion Models
Wentao Qu, Jing Wang, YongShun Gong, Xiaoshui Huang, Liang Xiao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27325-27335
Existing conditional Denoising Diffusion Probabilistic Models (DDPMs) with a Noise-Conditional Framework (NCF) remain challenging for 3D scene understanding tasks, as the complex geometric details in scenes increase the difficulty of fitting the gradients of the data distribution (the scores) from semantic labels. This also results in longer training and inference time for DDPMs compared to non-DDPMs. From a different perspective, we delve deeply into the model paradigm dominated by the Conditional Network. In this paper, we propose an end-to-end robust semantic Segmentation Network based on a Conditional-Noise Framework (CNF) of DDPMs, named CDSegNet. Specifically, CDSegNet models the Noise Network (NN) as a learnable noise-feature generator. This enables the Conditional Network (CN) to understand 3D scene semantics under multi-level feature perturbations, enhancing the generalization in unseen scenes. Meanwhile, benefiting from the noise system of DDPMs, CDSegNet exhibits strong robustness for data noise and sparsity in experiments. Moreover, thanks to CNF, CDSegNet can generate the semantic labels in a single-step inference like non-DDPMs, due to avoiding directly fitting the scores from semantic labels in the dominant network of CDSegNet. On public indoor and outdoor benchmarks, CDSegNet significantly outperforms existing methods, achieving state-of-the-art performance.
****

From Zero to Detail: Deconstructing Ultra-High-Definition Image Restoration from Progressive Spectral Perspective
Chen Zhao, Zhizhou Chen, Yunzhe Xu, Enxuan Gu, Jian Li, Zili Yi, Qian Wang, Jian Yang, Ying Tai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17935-17946

Ultra-high-definition (UHD) image restoration faces significant challenges due t
o its high resolution, complex content, and intricate details. To cope with thes
e challenges, we analyze the restoration process in depth through a progressive
spectral perspective, and deconstruct the complex UHD restoration problem into t
hree progressive stages: zero-frequency enhancement, low-frequency restoration,
and high-frequency refinement. Building on this insight, we propose a novel fram
ework, ERR, which comprises three collaborative sub-networks: the zero-frequenc
y enhancer (ZFE), the low-frequency restorer (LFR), and the high-frequency refin
er (HFR). Specifically, the ZFE integrates global priors to learn global mapping
, while the LFR restores low-frequency information, emphasizing reconstruction o
f coarse-grained content. Finally, the HFR employs our designed frequency-windo
wed Kolmogorov-Arnold Networks (FW-KAN) to refine textures and details, producin
g high-quality image restoration. Our approach significantly outperforms previou
s UHD methods across various tasks, with extensive ablation studies validating t
he effectiveness of each component.
****

SMILE: Infusing Spatial and Motion Semantics in Masked Video Learning
Fida Mohammad Thoker, Letian Jiang, Chen Zhao, Bernard Ghanem; Proceedings of th
e Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8438-8449
Masked video modeling, such as VideoMAE, is an effective paradigm for video self
-supervised learning (SSL). However, they are primarily based on reconstructing
pixel level details on natural videos which have substantial temporal redundancy
, limiting their capability for semantic representation and sufficient encoding
of motion dynamics. To address these issues, this paper introduces a novel SSL a
pproach for video representation learning, dubbed as SMILE, by infusing both spa
tial and motion semantics. In SMILE, we leverage image-language pretrained model
s, such as CLIP, to guide the learning process with their high-level spatial sem
antics. We enhance the representation of motion by introducing synthetic motion
patterns in the training data, allowing the model to capture more complex and dy
namic content. Furthermore, using SMILE, we establish a new self-supervised vide
o learning paradigm capable of learning strong video representations without req
uiring any natural video data. We have carried out extensive experiments on 7 da
tasets with various downstream scenarios. SMILE surpasses current state-of-the-a
rt SSL methods, showcasing its effectiveness in learning more discriminative and
generalizable video representations.
****

Video Language Model Pretraining with Spatio-temporal Masking
Yue Wu, Zhaobo Qi, Junshu Sun, Yaowei Wang, Qingming Huang, Shuhui Wang; Proceed
ings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp.
8557-8567
The development of self-supervised video-language models based on mask learning
has significantly advanced downstream video tasks. These models leverage masked
reconstruction to facilitate joint learning of visual and linguistic information
. However, recent study reveals that reconstructing image features yields superi
or downstream performance compared to video feature reconstruction. We hypothesi
ze that this performance gap stems from the way how masking strategies influence
the model's attention to temporal dynamics. To validate this hypothesis, we per
formed two sets of experiments that demonstrate that alignment between the maske
d target and the reconstruction target is crucial for self-supervised video-lang
uage learning. Based on these findings, we propose a spatio-temporal masking str
ategy (STM) for video-language model pretraining that operates across adjacent f
rames, and a decoder leverages semantic information to enhance the spatio-tempor
al representations of masked tokens. Thanks to the combination of masking strate
gy and reconstruction decoder, STM enforces the model to learn spatio-temporal f
eature representation comprehensively. Experiments in three video understanding
downstream tasks validate the superiority of our method.
****

COSMOS: Cross-Modality Self-Distillation for Vision Language Pre-training
Sanghwan Kim, Rui Xiao, Mariana-Iuliana Georgescu, Stephan Alaniz, Zeynep Akata;
Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2

025, pp. 14690-14700
Vision-Language Models (VLMs) trained with contrastive loss have achieved signif icant advancements in various vision and language tasks. However, the global nat ure of the contrastive loss makes VLMs focus predominantly on foreground objects , neglecting other crucial information in the image, which limits their effectiv eness in downstream tasks. To address these challenges, we propose COSMOS: CrOSs -MOdality Self-distillation for vision-language pre-training that integrates a n ovel text-cropping strategy and cross-attention module into a self-supervised le arning framework. We create global and local views of images and texts (i.e., mu lti-modal augmentations), which are essential for self-distillation in VLMs. We further introduce a cross-attention module, enabling COSMOS to learn comprehensi ve cross-modal representations optimized via a cross-modality self-distillation loss. COSMOS consistently outperforms previous strong baselines on various zero- shot downstream tasks, including retrieval, classification, and semantic segment ation. Additionally, it surpasses CLIP-based models trained on larger datasets i n visual perception and contextual understanding tasks. Code is available at htt ps://github.com/ExplainableML/cosmos.
****

Lifting Motion to the 3D World via 2D Diffusion
Jiaman Li, C. Karen Liu, Jiajun Wu; Proceedings of the Computer Vision and Patte rn Recognition Conference (CVPR), 2025, pp. 17518-17528
Estimating 3D motion from 2D observations is a long-standing research challenge. Prior work typically requires training on datasets containing ground truth 3D m otions, limiting their applicability to activities well-represented in existing motion capture data. This dependency particularly hinders generalization to out- of-distribution scenarios or subjects where collecting 3D ground truth is challe nging, such as complex athletic movements or animal motion. We introduce MVLift, a novel approach to predict global 3D motion---including both joint rotations a nd root trajectories in the world coordinate system---using only 2D pose sequenc es for training. Our multi-stage framework leverages 2D motion diffusion models to progressively generate consistent 2D pose sequences across multiple views, a key step in recovering accurate global 3D motion. MVLift generalizes across vari ous domains, including human poses, human-object interactions, and animal poses. Despite not requiring 3D supervision, it outperforms prior work on five dataset s, including those methods that require 3D supervision.
****

TAPT: Test-Time Adversarial Prompt Tuning for Robust Inference in Vision-Languag e Models
Xin Wang, Kai Chen, Jiaming Zhang, Jingjing Chen, Xingjun Ma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19910-1992 0
Large pre-trained Vision-Language Models (VLMs) such as CLIP have demonstrated e xcellent zero-shot generalizability across various downstream tasks. However, re cent studies have shown that the inference performance of CLIP can be greatly de graded by small adversarial perturbations, especially its visual modality, posin g significant safety threats. To mitigate this vulnerability, in this paper, we propose a novel defense method called Test-Time Adversarial Prompt Tuning (TAPT) to enhance the inference robustness of CLIP against visual adversarial attacks. TAPT is a test-time defense method that learns defensive bimodal (textual and v isual) prompts to robustify the inference process of CLIP. Specifically, it is a n unsupervised method that optimizes the defensive prompts for each test sample by minimizing a multi-view entropy and aligning adversarial-clean distributions. We evaluate the effectiveness of TAPT on 11 benchmark datasets, including Image Net and 10 other zero-shot datasets, demonstrating that it enhances the zero-sho t adversarial robustness of the original CLIP by at least 48.9% against AutoAtta ck (AA), while largely maintaining performance on clean examples. Moreover, TAPT outperforms existing adversarial prompt tuning methods across various backbones , achieving an average robustness improvement of at least 36.6%. Code is availab le at https://github.com/xinwong/TAPT.
****

Active Data Curation Effectively Distills Large-Scale Multimodal Models
Vishaal Udandarao, Nikhil Parthasarathy, Muhammad Ferjad Naeem, Talfan Evans, Samuel Albanie, Federico Tombari, Yongqin Xian, Alessio Tonioni, Olivier J. Henaff ; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14422-14437
Knowledge distillation (KD) is the de facto standard for compressing large-scale models into smaller ones. Prior works have explored ever more complex KD strategies involving different objective functions, teacher-ensembles, and weight inheritance. In this work we explore an alternative, yet simple approach---active data curation as effective distillation for contrastive multimodal pretraining. Our simple online batch selection method, ACID, outperforms strong KD baselines across various model-, data- and compute-configurations. Further, we find such an active data curation strategy to in fact be complementary to standard KD, and can be effectively combined to train highly performant inference-efficient models. Our simple and scalable pretraining framework, ACED, achieves state-of-the-art results across 27 zero-shot classification and retrieval tasks with upto 11% less inference FLOPs. We further demonstrate that our ACED models yield strong vision-encoders for training generative multimodal models in the LiT-Decoder setting, outperforming larger vision encoders for image-captioning and visual question-answering tasks.
****

PCDreamer: Point Cloud Completion Through Multi-view Diffusion Priors
Guangshun Wei, Yuan Feng, Long Ma, Chen Wang, Yuanfeng Zhou, Changjian Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27243-27253
This paper presents PCDreamer, a novel method for point cloud completion. Traditional methods typically extract features from partial point clouds to predict missing regions, but the large solution space often leads to unsatisfactory results. More recent approaches have started to use images as extra guidance, effectively improving performance, but obtaining paired data of images and partial point clouds is challenging in practice. To overcome these limitations, we harness the relatively view-consistent multi-view diffusion priors within large models, to generate novel views of the desired shape. The resulting image set encodes both global and local shape cues, which are especially beneficial for shape completion. To fully exploit the priors, we have designed a shape fusion module for producing an initial complete shape from multi-modality input (i.e., images and point clouds), and a follow-up shape consolidation module to obtain the final complete shape by discarding unreliable points introduced by the inconsistency from diffusion priors. Extensive experimental results demonstrate our superior performance, especially in recovering fine details.
****

Your ViT is Secretly an Image Segmentation Model
Tommie Kerssies, Niccolò Cavagnero, Alexander Hermans, Narges Norouzi, Giuseppe Averta, Bastian Leibe, Gijs Dubbelman, Daan de Geus; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25303-25313
Vision Transformers (ViTs) have shown remarkable performance and scalability across various computer vision tasks. To apply single-scale ViTs to image segmentation, existing methods adopt a convolutional adapter to generate multi-scale features, a pixel decoder to fuse these features, and a Transformer decoder that uses the fused features to make predictions. In this paper, we show that the inductive biases introduced by these task-specific components can instead be learned by the ViT itself, given sufficiently large models and extensive pre-training. Based on these findings, we introduce the Encoder-only Mask Transformer (EoMT), which repurposes the plain ViT architecture to conduct image segmentation. With large-scale models and pre-training, EoMT obtains a segmentation accuracy similar to state-of-the-art models that use task-specific components. At the same time, EoMT is significantly faster than these methods due to its architectural simplicity, e.g., up to 4x faster with ViT-L. Across a range of model sizes, EoMT demonstrates an optimal balance between segmentation accuracy and prediction speed, suggesting that compute resources are better spent on scaling the ViT itself rath

er than adding architectural complexity. Code: https://www.tue-mps.org/eomt/.
****
Cross-Rejective Open-Set SAR Image Registration
Shasha Mao, Shiming Lu, Zhaolong Du, Licheng Jiao, Shuiping Gou, Luntian Mou, Xu equan Lu, Lin Xiong, Yimeng Zhang; Proceedings of the Computer Vision and Patter n Recognition Conference (CVPR), 2025, pp. 23027-23036
Synthetic Aperture Radar (SAR) image registration is an essential upstream task in geoscience applications, in which pre-detected keypoints from two images are employed as observed objects to seek matched-point pairs. In general, the regist ration is regarded as a typical closed-set classification, which forces each key point to be classified into the given classes, but ignoring an essential issue t hat numerous redundant keypoints are beyond the given classes, which unavoidably results in capturing incorrect matched-point pairs. Based on this, we propose a Cross-Rejective Open-set SAR Image Registration (CroR-OSIR) method. In this wor k, these redundant keypoints are regarded as out-of-distribution (OOD) samples, and we formulate the registration as a special open-set task with two modules: s upervised contrastive feature-tuning and cross-rejective open-set recognition (C roR-OSR). Unlike traditional open-set recognition, all samples, including OOD sa mples, are available in the CroR-OSR module. CroR-OSR conducts the closed-set cl assifications in individual open-set domains from two images, meanwhile employin g the cross-domain rejection during training, to exclude these OOD samples based on confidence and consistency. Moreover, a new supervised contrastive tuning st rategy is incorporated for feature-tuning. Especially, the cross-domain estimati on labels obtained by CroR-OSR are fed back to the feature-tuning module for fea ture-tuning, to enhance feature discriminability. The experimental results illus trate that the proposed method achieves more precise registration than the state -of-the-art methods. The code is released at https://github.com/XDyaoshi/CroR-OS IR-main.
****
Synthetic Data is an Elegant GIFT for Continual Vision-Language Models
Bin Wu, Wuxuan Shi, Jinqiao Wang, Mang Ye; Proceedings of the Computer Vision an d Pattern Recognition Conference (CVPR), 2025, pp. 2813-2823
Pre-trained Vision-Language Models (VLMs) require Continual Learning (CL) to eff iciently update their knowledge and adapt to various downstream tasks without re training from scratch. However, for VLMs, in addition to the loss of knowledge p reviously learned from downstream tasks, pre-training knowledge is also corrupte d during continual fine-tuning. This issue is exacerbated by the unavailability of original pre-training data, leaving VLM's generalization ability degrading. I n this paper, we propose GIFT, a novel continual fine-tuning approach that utili zes synthetic data to overcome catastrophic forgetting in VLMs. Taking advantage of recent advances in text-to-image synthesis, we employ a pre-trained diffusio n model to recreate both pre-training and learned downstream task data. In this way, the VLM can revisit previous knowledge through distillation on matching dif fusion-generated images and corresponding text prompts. Leveraging the broad dis tribution and high alignment between synthetic image-text pairs in VLM's feature space, we propose a contrastive distillation loss along with an image-text alig nment constraint. To further combat in-distribution overfitting and enhance dist illation performance with limited amount of generated data, we incorporate adapt ive weight consolidation, utilizing Fisher information from these synthetic imag e-text pairs and achieving a better stability-plasticity balance. Extensive expe riments demonstrate that our method consistently outperforms previous state-of-t he-art approaches across various settings.
****
SplineGS: Robust Motion-Adaptive Spline for Real-Time Dynamic 3D Gaussians from Monocular Video
Jongmin Park, Minh-Quan Viet Bui, Juan Luis Gonzalez Bello, Jaeho Moon, Jihyong Oh, Munchurl Kim; Proceedings of the Computer Vision and Pattern Recognition Con ference (CVPR), 2025, pp. 26866-26875
Synthesizing novel views from in-the-wild monocular videos is challenging due to scene dynamics and the lack of multi-view cues. To address this, we propose Spl

ineGS, a COLMAP-free dynamic 3D Gaussian Splatting (3DGS) framework for high-qua lity reconstruction and fast rendering from monocular videos. At its core is a n ovel Motion-Adaptive Spline (MAS) method, which represents continuous dynamic 3D Gaussian trajectories using cubic Hermite splines with a small number of contro l points. For MAS, we introduce a Motion-Adaptive Control points Pruning (MACP) method to model the deformation of each dynamic 3D Gaussian across varying motio ns, progressively pruning control points while maintaining dynamic modeling inte grity. Additionally, we present a joint optimization strategy for camera paramet er estimation and 3D Gaussian attributes, leveraging photometric and geometric c onsistency. This eliminates the need for Structure-from-Motion preprocessing and enhances SplineGS's robustness in real-world conditions. Experiments show that SplineGS significantly outperforms state-of-the-art methods in novel view synthe sis quality for dynamic scenes from monocular videos, achieving thousands times faster rendering speed.

****

SCSA: A Plug-and-Play Semantic Continuous-Sparse Attention for Arbitrary Semanti c Style Transfer

Chunnan Shang, Zhizhong Wang, Hongwei Wang, Xiangming Meng; Proceedings of the C omputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13051-13060

Attention-based arbitrary style transfer methods, including CNN-based, Transform er-based, and Diffusion-based, have flourished and produced high-quality stylize d images. However, they perform poorly on the content and style images with the same semantics, i.e., the style of the corresponding semantic region of the gene rated stylized image is inconsistent with that of the style image. We argue that the root cause lies in their failure to consider the relationship between local regions and semantic regions. To address this issue, we propose a plug-and-play semantic continuous-sparse attention, dubbed SCSA, for arbitrary semantic style transfer---each query point considers certain key points in the corresponding s emantic region. Specifically, semantic continuous attention ensures each query p oint fully attends to all the continuous key points in the same semantic region that reflect the overall style characteristics of that region; Semantic sparse a ttention allows each query point to focus on the most similar sparse key point i n the same semantic region that exhibits the specific stylistic texture of that region. By combining the two modules, the resulting SCSA aligns the overall styl e of the corresponding semantic regions while transferring the vivid textures of these regions. Qualitative and quantitative results demonstrate that SCSA enabl es attention-based arbitrary style transfer methods to produce high-quality sema ntic stylized images.

****

Timestep Embedding Tells: It's Time to Cache for Video Diffusion Model

Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Ying ya Zhang, Qixiang Ye, Fang Wan; Proceedings of the Computer Vision and Pattern R ecognition Conference (CVPR), 2025, pp. 7353-7363

As a fundamental backbone for video generation, diffusion models are challenged by low inference speed due to the sequential nature of denoising.Previous method s speed up the models by caching and reusing model outputs at uniformly selected timesteps.However, such a strategy neglects the fact that differences among mod el outputs are not uniform across timesteps, which hinders selecting the appropr iate model outputs to cache, leading to a poor balance between inference efficie ncy and visual quality.In this study, we introduce Timestep Embedding Aware Cach e (TeaCache), a training-free caching approach that estimates and leverages the fluctuating differences among model outputs across timesteps.Rather than directl y using the time-consuming model outputs, TeaCache focuses on model inputs, whic h have a strong correlation with the modeloutputs while incurring negligible com putational cost.TeaCache first modulates the noisy inputs using the timestep emb eddings to ensure their differences better approximating those of model outputs. TeaCache then introduces a rescaling strategy to refine the estimated differenc es and utilizes them to indicate output caching.Experiments show that TeaCache a chieves up to 4.41xacceleration over Open-Sora-Plan with negligible (-0.07% Vben ch score) degradation of visual quality.

****
Can't Slow Me Down: Learning Robust and Hardware-Adaptive Object Detectors against Latency Attacks for Edge Devices
Tianyi Wang, Zichen Wang, Cong Wang, Yuanchao Shu, Ruilong Deng, Peng Cheng, Jiming Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19230-19240
Object detection is a fundamental enabler for many real-time downstream applications such as autonomous driving, augmented reality and supply chain management. However, the algorithmic backbone of neural networks is brittle to imperceptible perturbations in the system inputs, which were generally known as misclassifying attacks. By targeting the real-time processing capability, a new class of latency attacks has been reported recently. They exploit new attack surfaces in object detectors by creating a computational bottleneck in the post-processing module, which leads to cascading failure and puts the real-time downstream tasks at risk. In this work, we take an initial attempt to defend against this attack via background-attentive adversarial training that is also cognizant of the underlying hardware capabilities. We first draw system-level connections between latency attacks and hardware capacity across heterogeneous GPU devices. Based on the particular adversarial behaviors, we utilize objectness loss as a proxy and build background attention into the adversarial training pipeline, and achieve a favorable balance between clean and robust accuracy. The extensive experiments demonstrate the effectiveness of the defense in restoring real-time processing capability from 13 FPS to 43 FPS on Jetson Orin NX, with a better trade-off between the clean and robust accuracy. The source code is available at: https://github.com/Hill-Wu-1998/underload.
****
Multi-modal Knowledge Distillation-based Human Trajectory Forecasting
Jaewoo Jeong, Seohee Lee, Daehee Park, Giwon Lee, Kuk-Jin Yoon; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24222-24233
Pedestrian trajectory forecasting is crucial in various applications such as autonomous driving and mobile robot navigation. In such applications, camera-based perception enables the extraction of additional modalities (human pose, text) to enhance prediction accuracy. Indeed, we find that textual descriptions play a crucial role in integrating additional modalities into a unified understanding. However, online extraction of text requires the use of VLM, which may not be feasible for resource-constrained systems. To address this challenge, we propose a multi-modal knowledge distillation framework: a student model with limited modality is distilled from a teacher model trained with full range of modalities. The comprehensive knowledge of a teacher model trained with trajectory, human pose, and text is distilled into a student model using only trajectory or human pose as a sole supplement. In doing so, we separately distill the core locomotion insights from intra-agent multi-modality and inter-agent interaction. Our generalizable framework is validated with two state-of-the-art models across three datasets on both ego-view (JRDB, SIT) and BEV-view (ETH/UCY) setups, utilizing both annotated and VLM-generated text captions. Distilled student models show consistent improvement in all prediction metrics for both full and instantaneous observations, improving up to 13%. The code is available at github.com/Jaewoo97/KDTF.
****
SAM2Object: Consolidating View Consistency via SAM2 for Zero-Shot 3D Instance Segmentation
Jihuai Zhao, Junbao Zhuo, Jiansheng Chen, Huimin Ma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19325-19334
In the field of zero-shot 3D instance segmentation, existing 2D-to-3D lifting methods typically obtain 2D segmentation across multiple RGB frames using vision foundation models, which are then projected and merged into 3D space. However, since the inference of vision foundation models on a single frame is not integrated with adjacent frames, the masks of the same object may vary across different frames, leading to a lack of view consistency in the 2D segmentation. Furthermore, current lifting methods average the 2D segmentation from multiple views during

the projection into 3D space, causing low-quality masks and high-quality masks to share the same weight. These factors can lead to fragmented 3D segmentation. In this paper, we present SAM2Object, a novel zero-shot 3D instance segmentation method that effectively utilizes the Segment Anything Model 2 to segment and track objects, consolidating view consistency across frames. Our approach combines these consistent 2D masks with 3D geometric priors, improving the robustness of 3D segmentation. Additionally, we introduce mask consolidation module to filter out low-quality masks across frames, which enables more precise 2D-to-3D matching. Comprehensive evaluations on ScanNetV2, ScanNet++ and ScanNet200 demonstrate the robustness and effectiveness of SAM2Object, showcasing its ability to outperform previous methods. Our project page is at https://jihuaizhaohd.github.io/SAM2Object.

****

CDI: Copyrighted Data Identification in Diffusion Models

Jan Dubi■ski, Antoni Kowalczuk, Franziska Boenisch, Adam Dziedzic; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18674-18684

Diffusion Models (DMs) benefit from large and diverse datasets for their training. Since this data is often scraped from the Internet without permission from the data owners, this raises concerns about copyright and intellectual property protections. While (illicit) use of data is easily detected for training samples perfectly re-created by a DM at inference time, it is much harder for data owners to verify if their data was used for training when the outputs from the suspect DM are not close replicas. Conceptually, membership inference attacks (MIAs), which detect if a given data point was used during training, present themselves as a suitable tool to address this challenge. However, we demonstrate that existing MIAs are not strong enough to reliably determine the membership of individual images in large, state-of-the-art DMs. To overcome this limitation, we propose Copyrighted Data Identification (CDI), a framework for data owners to identify whether their dataset was used to train a given DM. CDI relies on dataset inference techniques, i.e., instead of using the membership signal from a single data point, CDI leverages the fact that most data owners, such as providers of stock photography, visual media companies, or even individual artists, own datasets with multiple publicly exposed data points which might all be included in the training of a given DM. By selectively aggregating signals from existing MIAs and using new handcrafted methods to extract features from these datasets, feeding them to a scoring model, and applying rigorous statistical testing, CDI allows data owners with as little as 70 data points to identify with a confidence of more than 99% whether their data was used to train a given DM. Thereby, CDI represents a valuable tool for data owners to claim illegitimate use of their copyrighted data.

****

Hypergraph Vision Transformers: Images are More than Nodes, More than Edges

Joshua Fixelle; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9751-9761

Recent advancements in computer vision have highlighted the scalability of Vision Transformers (ViTs) across various tasks, yet challenges remain in balancing adaptability, computational efficiency, and the ability to model higher-order relationships. Vision Graph Neural Networks (ViGs) offer an alternative by leveraging graph-based methodologies but are hindered by the computational bottlenecks of clustering algorithms used for edge generation. To address these issues, we propose the Hypergraph Vision Transformer (HgVT), which incorporates a hierarchical bipartite hypergraph structure into the vision transformer framework to capture higher-order semantic relationships while maintaining computational efficiency. HgVT leverages population and diversity regularization for dynamic hypergraph construction without clustering, and expert edge pooling to enhance semantic extraction and facilitate graph-based image retrieval. Empirical results demonstrate that HgVT achieves strong performance on image classification and retrieval, positioning it as an efficient framework for semantic-based vision tasks.

****

Binarized Neural Network for Multi-spectral Image Fusion
Junming Hou, Xiaoyu Chen, Ran Ran, Xiaofeng Cong, Xinyang Liu, Jian Wei You, Liang-Jian Deng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2236-2245
Pan-sharpening technology refers to generating a high-resolution (HR) multi-spectral (MS) image with broad applications by fusing a low-resolution (LR) MS image and HR panchromatic (PAN) image. While deep learning approaches have shown impressive performance in pan-sharpening, they generally require extensive hardware with high memory and computational power, limiting their deployment on resource-constrained satellites. In this study, we investigate the use of binary neural networks (BNNs) for pan-sharpening and observe that binarization leads to distinct information degradation across different frequency components of an image. Building on this insight, we propose a novel binary pan-sharpening network, termed BNNPan, structured around the Prior-Integrated Binary Frequency (PIBF) module that features three key ingredients: Binary Wavelet Transform Convolution, Latent Diffusion Prior Compensation, and Channel-wise Distribution Calibration. Specifically, the first decomposes input features into distinct frequency components using Wavelet Transform, then applies a "divide-and-conquer" strategy to optimize binary feature learning for each component, informed by the corresponding full-precision residual statistics. The second integrates a latent diffusion prior to compensate for compromised information during binarization, while the third performs channel-wise calibration to further refine feature representation. Our BNNPan, developed with the proposed techniques, achieves promising pan-sharpening performance on multiple remote sensing datasets, surpassing state-of-the-art binarization algorithms.
****
CRISP: Object Pose and Shape Estimation with Test-Time Adaptation
Jingnan Shi, Rajat Talak, Harry Zhang, David Jin, Luca Carlone; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11644-11653
We consider the problem of estimating object pose and shape from an RGB-D image. Our first contribution is to introduce CRISP, a category-agnostic object pose and shape estimation pipeline. The pipeline implements an encoder-decoder model for shape estimation. It uses FiLM-conditioning for implicit shape reconstruction and a DPT-based network for estimating pose-normalized points for pose estimation. As a second contribution, we propose an optimization-based pose and shape corrector that can correct estimation errors caused by a domain gap. Observing that the shape decoder is well behaved in the convex hull of known shapes, we approximate the shape decoder with an active shape model, and show that this reduces the shape correction problem to a constrained linear least squares problem, which can be solved efficiently by an interior point algorithm. Third, we introduce a self-training pipeline to perform self-supervised domain adaptation of CRISP. The self-training is based on a correct-and-certify approach, which leverages the corrector to generate pseudo-labels at test time, and uses them to self-train CRISP. We demonstrate CRISP (and the self-training) on YCBV, SPE3R, and NOCS datasets. CRISP shows high performance on all the datasets. Moreover, our self-training is capable of bridging a large domain gap. Finally, CRISP also shows an ability to generalize to unseen objects. Code, pre-trained models and videos of sample results are available on the project webpage https://web.mit.edu/sparklab/research/crisp_object_pose_shape/.
****
ShiftwiseConv: Small Convolutional Kernel with Large Kernel Effect
Dachong Li, Li Li, Zhuangzhuang Chen, Jianqiang Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25281-25291
Large kernels make standard convolutional neural networks (CNNs) great again over transformer architectures in various vision tasks. Nonetheless, recent studies meticulously designed around increasing kernel size have shown diminishing returns or stagnation in performance. Thus, the hidden factors of large kernel convolution that affect model performance remain unexplored. In this paper, we reveal that the key hidden factors of large kernels can be summarized as two separate

components: extracting features at a certain granularity and fusing features by multiple pathways. To this end, we leverage the multi-path long-distance sparse dependency relationship to enhance feature utilization via the proposed Shiftwise (SW) convolution operator with a pure CNN architecture. In a wide range of vision tasks such as classification, segmentation, and detection, SW surpasses state-of-the-art transformers and CNN architectures, including SLaK and UniRepLKNet. More importantly, our experiments demonstrate that 3 x3 convolutions can replace large convolutions in existing large kernel CNNs to achieve comparable effects, which may inspire follow-up works. Code and all the models at https://github.com/lidc54/shift-wiseConv.

****

GaussianIP: Identity-Preserving Realistic 3D Human Generation via Human-Centric Diffusion Prior

Zichen Tang, Yuan Yao, Miaomiao Cui, Liefeng Bo, Hongyu Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 348-358

Text-guided 3D human generation has advanced with the development of efficient 3D representations and 2D-lifting methods like score distillation sampling (SDS). However, current methods suffer from prolonged training times and often produce results that lack fine facial and garment details. In this paper, we propose GaussianIP, an effective two-stage framework for generating identity-preserving realistic 3D humans from text and image prompts. Our core insight is to leverage human-centric knowledge to facilitate the generation process. In stage 1, we propose a novel Adaptive Human Distillation Sampling (AHDS) method to rapidly generate a 3D human that maintains high identity consistency with the image prompt and achieves a realistic appearance. Compared to traditional SDS methods, AHDS better aligns with the human-centric generation process, enhancing visual quality with notably fewer training steps. To further improve the visual quality of the face and clothes regions, we design a View-Consistent Refinement (VCR) strategy in stage 2. Specifically, it produces detail-enhanced results of the multi-view images from stage 1 iteratively, ensuring the 3D texture consistency across views via mutual attention and distance-guided attention fusion. Then a polished version of the 3D human can be achieved by directly perform reconstruction with the refined images. Extensive experiments demonstrate that GaussianIP outperforms existing methods in both visual quality and training efficiency, particularly in generating identity-preserving results.

****

Creating Your Editable 3D Photorealistic Avatar with Tetrahedron-constrained Gaussian Splatting

Hanxi Liu, Yifang Men, Zhouhui Lian; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15976-15986

Personalized 3D avatar editing holds significant promise due to its user-friendliness and availability to applications such as AR/VR and virtual try-ons. Previous studies have explored the feasibility of 3D editing, but often struggle to generate visually pleasing results, possibly due to the unstable representation learning under mixed optimization of geometry and texture in complicated reconstructed scenarios. In this paper, we aim to provide an accessible solution for ordinary users to create their editable 3D avatars with precise region localization, geometric adaptability, and photorealistic renderings. To tackle this challenge, we introduce a meticulously designed framework that decouples the editing process into local spatial adaptation and realistic appearance learning, utilizing a hybrid Tetrahedron-constrained Gaussian Splatting (TetGS) as the underlying representation. TetGS combines the controllable explicit structure of tetrahedral grids with the high-precision rendering capabilities of 3D Gaussian Splatting and is optimized in a progressive manner comprising three stages: 3D avatar instantiation from real-world monocular videos to provide accurate priors for TetGS initialization; localized spatial adaptation with explicitly partitioned tetrahedrons to guide the redistribution of Gaussian kernels; and geometry-based appearance generation with a coarse-to-fine activation strategy. Both qualitative and quantitative experiments demonstrate the effectiveness and superiority of our approach in generating photorealistic 3D editable avatars.

****
FineVQ: Fine-Grained User Generated Content Video Quality Assessment
Huiyu Duan, Qiang Hu, Jiarui Wang, Liu Yang, Zitong Xu, Lu Liu, Xiongkuo Min, Chunlei Cai, Tianxiao Ye, Xiaoyun Zhang, Guangtao Zhai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3206-3217
The rapid growth of user-generated content (UGC) videos has produced an urgent need for effective video quality assessment (VQA) algorithms to monitor video quality and guide optimization and recommendation procedures. However, current VQA models generally only give an overall rating for a UGC video, which lacks fine-grained labels for serving video processing and recommendation applications. To address the challenges and promote the development of UGC videos, we establish the first large-scale Fine-grained Video quality assessment Database, termed FineVD, which comprises 6104 UGC videos with fine-grained quality scores and descriptions across multiple dimensions. Based on this database, we propose a Fine-grained Video Quality assessment (FineVQ) model to learn the fine-grained quality of UGC videos, with the capabilities of quality rating, quality scoring, and quality attribution. Extensive experimental results demonstrate that our proposed FineVQ can produce fine-grained video-quality results and achieve state-of-the-art performance on FineVD and other commonly used UGC-VQA datasets. Both FineVD and FineVQ are publicly available at: https://github.com/IntMeGroup/FineVQ.
****
Unveiling the Ignorance of MLLMs: Seeing Clearly, Answering Incorrectly
Yexin Liu, Zhengyang Liang, Yueze Wang, Xianfeng Wu, Feilong Tang, Muyang He, Jian Li, Zheng Liu, Harry Yang, Sernam Lim, Bo Zhao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9087-9097
Multimodal **L**arge **L**anguage **M**odels (MLLMs) have displayed remarkable performance in multimodal tasks, particularly in visual comprehension. However, we reveal that MLLMs often generate incorrect answers even when they understand the visual content. To this end, we manually construct a benchmark with 12 categories and design evaluation metrics that assess the degree of error in MLLM responses even when the visual content is seemingly understood. Based on this benchmark, we test 15 leading MLLMs and analyze the distribution of attention maps and logits of some MLLMs. Our investigation identifies two primary issues: 1) most instruction tuning datasets predominantly feature questions that "directly" relate to the visual content, leading to a bias in MLLMs' responses to other indirect questions, and 2) MLLMs' attention to visual tokens is notably lower than to system and question tokens. We further observe that attention scores between questions and visual tokens as well as the model's confidence in the answers are lower in response to misleading questions than to straightforward ones. To address the first challenge, we introduce a paired positive and negative data construction pipeline to diversify the dataset. For the second challenge, we propose to enhance the model's focus on visual content during decoding by refining the text and visual prompt. For the text prompt, we propose a content-guided refinement strategy that performs preliminary visual content analysis to generate structured information before answering the question. Additionally, we employ a visual attention refinement strategy that highlights question-relevant visual tokens to increase the model's attention to visual content that aligns with the question. Extensive experiments demonstrate that these challenges can be significantly mitigated with our proposed dataset and techniques. The benchmark, training set, and code will be available.
****
Object-Shot Enhanced Grounding Network for Egocentric Video
Yisen Feng, Haoyu Zhang, Meng Liu, Weili Guan, Liqiang Nie; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24190-24200
Egocentric video grounding is a crucial task for embodied intelligence applications, distinct from exocentric video moment localization. Existing methods primarily focus on the distributional differences between egocentric and exocentric videos but often neglect key characteristics of egocentric videos and the fine-grained information emphasized by question-type queries. To address these limitations, we propose OSGNet, an Object-Shot enhanced Grounding Network for egocentric

video. Specifically, we extract object information from videos to enrich video r
epresentation, particularly for objects highlighted in the textual query but not
 directly captured in the video features. Additionally, we analyze the frequent
shot movements inherent to egocentric videos, leveraging these features to extra
ct the wearer's attention information, which enhances the model's ability to per
form modality alignment. Experiments conducted on three datasets demonstrate tha
t OSGNet achieves state-of-the-art performance, validating the effectiveness of
our approach. Our code can be found at https://github.com/Yisen-Feng/OSGNet.
****

MANTA: Diffusion Mamba for Efficient and Effective Stochastic Long-Term Dense Ac
tion Anticipation

Olga Zatsarynna, Emad Bahrami, Yazan Abu Farha, Gianpiero Francesca, Juergen Gal
l; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR),
 2025, pp. 3438-3448

Long-term dense action anticipation is very challenging since it requires predic
ting actions and their durations several minutes into the future based on provid
ed video observations. To model the uncertainty of future outcomes, stochastic m
odels predict several potential future action sequences for the same observation
. Recent work has further proposed to incorporate uncertainty modelling for obse
rved frames by simultaneously predicting per-frame past and future actions in a
unified manner. While such joint modelling of actions is beneficial, it requires
 long-range temporal capabilities to connect events across distant past and futu
re time points. However, the previous work struggles to achieve such a long-rang
e understanding due to its limited and/or sparse receptive field. To alleviate t
his issue, we propose a novel MANTA (MAmba for ANTicipation) network. Our model
enables effective long-term temporal modelling even for very long sequences whil
e maintaining linear complexity in sequence length. We demonstrate that our appr
oach achieves state-of-the-art results on three datasets - Breakfast, 50Salads,
and Assembly101 - while also significantly improving computational and memory ef
ficiency. Our code is available at https://github.com/olga-zats/DIFF_MANTA.
****

METASCENES: Towards Automated Replica Creation for Real-world 3D Scans

Huangyue Yu, Baoxiong Jia, Yixin Chen, Yandan Yang, Puhao Li, Rongpeng Su, Jiaxi
n Li, Qing Li, Wei Liang, Song-Chun Zhu, Tengyu Liu, Siyuan Huang; Proceedings o
f the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1667-
1679

Embodied AI (EAI) research requires high-quality, diverse 3D scenes to effective
ly support skill acquisition, sim-to-real transfer, and generalization. Achievin
g these quality standards, however, necessitates the precise replication of real
-world object diversity. Existing datasets demonstrate that this process heavily
 relies on artist-driven designs, which demand substantial human effort and pres
ent significant scalability challenges. To scalably produce realistic and intera
ctive 3D scenes, we first present MetaScenes, a large-scale simulatable 3D scene
 dataset constructed from real-world scans, which includes 15366 objects spannin
g 831 fine-grained categories. Then, we introduce Scan2Sim, a robust multi-modal
 alignment model, which enables the automated, high-quality replacement of asset
s, thereby eliminating the reliance on artist-driven designs for scaling 3D scen
es. We further propose two benchmarks to evaluate MetaScenes: a detailed scene s
ynthesis task focused on small item layouts for robotic manipulation and a domai
n transfer task in vision-and-language navigation (VLN) to validate cross-domain
 transfer. Results confirm MetaScenes's potential to enhance EAI by supporting m
ore generalizable agent learning and sim-to-real applications, introducing new p
ossibilities for EAI research.
****

Robust Multimodal Survival Prediction with Conditional Latent Differentiation Va
riational AutoEncoder

Junjie Zhou, Jiao Tang, Yingli Zuo, Peng Wan, Daoqiang Zhang, Wei Shao; Proceedi
ngs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp.
10384-10393

The integrative analysis of histopathological images and genomic data has receiv

ed increasing attention for survival prediction of human cancers. However, the existing studies always hold the assumption that full modalities are available. As a matter of fact, the cost for collecting genomic data is high, which sometimes makes genomic data unavailable in testing samples. A common way of tackling such incompleteness is to generate the genomic representations from the pathology images. Nevertheless, such strategy still faces the following two challenges: (1) The gigapixel whole slide images (WSIs) are huge and thus hard for representation. (2) It is difficult to generate the genomic embeddings with diverse function categories in a unified generative framework. To address the above challenges, we propose a Conditional Latent Differentiation Variational AutoEncoder (LD-CVAE) for robust multimodal survival prediction, even with missing genomic data. Specifically, a Variational Information Bottleneck Transformer (VIB-Trans) module is proposed to learn compressed pathological representations from the gigapixel WSIs. To generate different functional genomic features, we develop a novel Latent Differentiation Variational AutoEncoder (LD-VAE) to learn the genomic and function-specific posteriors for the genomic embeddings with diverse functions. Finally, we use the product-of-experts technique to integrate the genomic posterior and image posterior for the joint latent distribution estimation in LD-CVAE. We test the effectiveness of our method on five different cancer datasets, and the experimental results demonstrate its superiority in both complete and missing modality scenarios.

****

Sim-to-Real Causal Transfer: A Metric Learning Approach to Causally-Aware Interaction Representations

Ahmad Rahimi, Po-Chien Luan, Yuejiang Liu, Frano Raji■, Alexandre Alahi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17271-17281

Modeling spatial-temporal interactions among neighboring agents is at the heart of multi-agent problems such as motion forecasting and crowd navigation. Despite notable progress, it remains unclear to which extent modern representations can capture the causal relationships behind agent interactions. In this work, we take an in-depth look at the causal awareness of these representations, from computational formalism to real-world practice. First, we revisit the notion of non-causal robustness studied in the recent CausalAgents benchmark. We show that existing representations are already partially resilient to perturbations of non-causal agents, and yet modeling indirect causal effects involving mediator agents remains challenging. To address this challenge, we introduce a metric learning approach that regularizes latent representations with causal annotations. Our controlled experiments show that this approach not only leads to higher degrees of causal awareness but also yields stronger out-of-distribution robustness. To further operationalize it in practice, we propose a sim-to-real causal transfer method via cross-domain multi-task learning. Experiments on trajectory prediction datasets show that our method can significantly boost generalization, even in the absence of real-world causal annotations, where we acquire higher prediction accuracy by only using 25% of real-world data. We hope our work provides a new perspective on the challenges and potential pathways toward causally-aware representations of multi-agent interactions. Our code is available in supplementary materials. Our code is available at https://github.com/vita-epfl/CausalSim2Real.

****

Ev-3DOD: Pushing the Temporal Boundaries of 3D Object Detection with Event Cameras

Hoonhee Cho, Jae-Young Kang, Youngho Kim, Kuk-Jin Yoon; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27197-27210

Detecting 3D objects in point clouds plays a crucial role in autonomous driving systems. Recently, advanced multi-modal methods incorporating camera information have achieved notable performance. For a safe and effective autonomous driving system, algorithms that excel not only in accuracy but also in speed and low latency are essential. However, existing algorithms fail to meet these requirements due to the latency and bandwidth limitations of fixed frame rate sensors, e.g., LiDAR and camera. To address this limitation, we introduce asynchronous event c

ameras into 3D object detection for the first time. We leverage their high tempo ral resolution and low bandwidth to enable high-speed 3D object detection. Our m ethod enables detection even during inter-frame intervals when synchronized data is unavailable, by retrieving previous 3D information through the event camera. Furthermore, we introduce the first event-based 3D object detection datasets, E v-Waymo and DSEC-3DOD, both of which include ground truth 3D bounding boxes at 1 00 FPS, establishing the first benchmarks for event-based 3D detectors. The code and dataset are available at https://github.com/mickeykang16/Ev3DOD.
****

Zero-Shot Blind-spot Image Denoising via Implicit Neural Sampling
Yuhui Quan, Tianxiang Zheng, Zhiyuan Ma, Hui Ji; Proceedings of the Computer Vis ion and Pattern Recognition Conference (CVPR), 2025, pp. 7502-7512
The blind-spot principle has been a widely used tool in zero-shot image denoisin g but faces challenges with real-world noise that exhibits strong local correlat ions. Existing methods focus on reducing noise correlation, which also weaken th e pixel correlations needed for accurately estimating missing pixels. In this pa per, we first present a rigorous analysis of how noise correlation and pixel cor relation impact the statistical risk of a linear blind-spot denoiser. We then pr opose using an implicit neural representation to resample noisy pixels, effectiv ely reducing noise correlation while preserving the essential pixel correlations for successful blind-spot denoising. Extensive experiments show our method surp asses existing zero-shot denoising techniques on real-world noisy images.
****

Nearly Zero-Cost Protection Against Mimicry by Personalized Diffusion Models
Namhyuk Ahn, KiYoon Yoo, Wonhyuk Ahn, Daesik Kim, Seung-Hun Nam; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28801-2 8810
Recent advancements in diffusion models revolutionize image generation but pose risks of misuse, such as replicating artworks or generating deepfakes. Existing image protection methods, though effective, struggle to balance protection effic acy, invisibility, and latency, thus limiting practical use. We introduce pertur bation pre-training to reduce latency and propose a mixture-of-perturbations app roach that dynamically adapts to input images to minimize performance degradatio n. Our novel training strategy computes protection loss across multiple VAE feat ure spaces, while adaptive targeted protection at inference enhances robustness and invisibility. Experiments show comparable protection performance with improv ed invisibility and drastically reduced inference time.
****

Tripartite Weight-Space Ensemble for Few-Shot Class-Incremental Learning
Juntae Lee, Munawar Hayat, Sungrack Yun; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15329-15338
Few-shot class incremental learning (FSCIL) enables the continual learning of ne w concepts with only a few training examples. In FSCIL, the model undergoes subs tantial updates, making it prone to forgetting previous concepts and overfitting to the limited new examples. Most recent trend is typically to disentangle the learning of the representation from the classification head of the model. A well -generalized feature extractor on the base classes (many examples and many class es) is learned, and then fixed during incremental learning. Arguing that the fix ed feature extractor restricts the model's adaptability to new classes, we intro duce a novel FSCIL method to effectively address catastrophic forgetting and ove rfitting issues. Our method enables to seamlessly update the entire model with a few examples. We mainly propose a tripartite weight-space ensemble (Tri-WE). Tr i-WE interpolates the base, immediately previous, and current models in weight-s pace, especially for the classification heads of the models. Then, it collaborat ively maintains knowledge from the base and previous models. In addition, we rec ognize the challenges of distilling generalized representations from the previou s model from scarce data. Hence, we suggest a regularization loss term using amp lified data knowledge distillation. Simply intermixing the few-shot data, we can produce richer data enabling the distillation of critical knowledge from the pr evious model. Consequently, we attain state-of-the-art results on the miniImageN

et, CUB200, and CIFAR100 datasets.
****
The Devil is in Temporal Token: High Quality Video Reasoning Segmentation
Sitong Gong, Yunzhi Zhuge, Lu Zhang, Zongxin Yang, Pingping Zhang, Huchuan Lu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29183-29192
Existing methods for Video Reasoning Segmentation rely heavily on a single special token to represent the object in the keyframe or the entire video, inadequately capturing spatial complexity and inter-frame motion. To overcome these challenges, we propose VRS-HQ, an end-to-end video reasoning segmentation approach that leverages Multimodal Large Language Models (MLLMs) to inject rich spatiotemporal features into hierarchical tokens. Our key innovations include a Temporal Dynamic Aggregation (TDA) and a Token-driven Keyframe Selection (TKS). Specifically, we design frame-level <SEG> and temporal-level <TAK> tokens that utilize MLLM's autoregressive learning to effectively capture both local and global information. Subsequently, we apply a similarity-based weighted fusion and frame selection strategy, then utilize SAM2 to perform keyframe segmentation and propagation. To enhance keyframe localization accuracy, the TKS filters keyframes based on SAM2's occlusion scores during inference. VRS-HQ achieves state-of-the-art performance on ReVOS, surpassing VISA by 5.9%/12.5%/9.1% in J&F scores across the three subsets. These results highlight the strong temporal reasoning and segmentation capabilities of our method.
****
PerLA: Perceptive 3D Language Assistant
Guofeng Mei, Wei Lin, Luigi Riz, Yujiao Wu, Fabio Poiesi, Yiming Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14369-14379
Enabling Large Language Models (LLMs) to understand the 3D physical world is an emerging yet challenging research direction. Current strategies for processing point clouds typically downsample the scene or divide it into smaller parts for separate analysis. However, both approaches risk losing key local details or global contextual information. In this paper, we introduce PerLA, a 3D language assistant designed to be more perceptive to both details and context, making visual representations more informative for the LLM. PerLA captures high-resolution (local) details in parallel from different point cloud areas and integrates them with (global) context obtained from a lower-resolution whole point cloud. We present a novel algorithm that preserves point cloud locality through the Hilbert curve and effectively aggregates local-to-global information via cross-attention and a graph neural network. Lastly, we introduce a novel loss for local representation consensus to promote training stability. PerLA outperforms state-of-the-art 3D language assistants, with gains of up to +1.34 CiDEr on ScanQA for question answering, and +4.22 on ScanRefer and +3.88 on Nr3D for dense captioning.
****
LITA-GS: Illumination-Agnostic Novel View Synthesis via Reference-Free 3D Gaussian Splatting and Physical Priors
Han Zhou, Wei Dong, Jun Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21580-21589
Directly employing 3D Gaussian Splatting (3DGS) on images with adverse illumination conditions exhibits considerable difficulty in achieving high-quality normally-exposed representation due to: (1) The limited Structure from Motion (SfM) points estimated in adverse illumination scenarios fail to capture sufficient scene details; (2) Without ground-truth references, the intensive information loss, huge noise, and color distortion poses substantial challenges for 3DGS to produce high-quality results; (3) Combining existing exposure correction methods with 3DGS can not achieve satisfactory performance due to their individual enhancement process, which leads to the illumination inconsistency between enhanced images from different viewpoints. To address these issues, we propose LITA-GS, a novel illumination-agnostic novel view synthesis method via reference-free 3DGS and physical priors. Firstly, we introduce an illumination-invariant physical prior extraction pipeline. Secondly, based on the extracted robust spatial structure pr

ior, we develop the lighting-agnostic structure rendering strategy, which facili tates the optimization of the scene structure and object appearance. Moreover, a progressive denoising module is introduced to effectively surpass the noise wit hin the light-invariant representation. We adopt the unsupervised strategy for t he training of LITA-GS and extensive experiments demonstrate that LITA-GS surpas s the state-of-the-art (SOTA) NeRF-based method by 1.7 dB in PSNR and 0.09 in SS IM while enjoying faster inference speed and costing reduced training time. The code will be released upon acceptance.

****

PhyT2V: LLM-Guided Iterative Self-Refinement for Physics-Grounded Text-to-Video Generation

Qiyao Xue, Xiangyu Yin, Boyuan Yang, Wei Gao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18826-18836

Text-to-video (T2V) generation has been recently enabled by transformer-based di ffusion models, but current T2V models lack capabilities in adhering to the real -world common knowledge and physical rules, due to their limited understanding o f physical realism and deficiency in temporal modeling. Existing solutions are e ither data-driven or require extra model inputs, but cannot be generalizable to out-of-distribution domains. In this paper, we present PhyT2V, a new data-indepe ndent T2V technique that expands the current T2V model's capability of video gen eration to out-of-distribution domains, by enabling chain-of-thought and step-ba ck reasoning in T2V prompting. Our experiments show that PhyT2V improves existin g T2V models' adherence to real-world physical rules by 2.3x, and achieves 35% i mprovement compared to T2V prompt enhancers.

****

Track4Gen: Teaching Video Diffusion Models to Track Points Improves Video Genera tion

Hyeonho Jeong, Chun-Hao P. Huang, Jong Chul Ye, Niloy J. Mitra, Duygu Ceylan; Pr oceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025 , pp. 7276-7287

While recent foundational video generators produce visually rich output, they st ill struggle with appearance drift, where objects gradually degrade or change in consistently across frames, breaking visual coherence. We hypothesize that this is because there is no explicit supervision in terms of spatial tracking at the feature level. We propose Track4Gen, a spatially aware video generator that comb ines video diffusion loss with point tracking across frames, providing enhanced spatial supervision on the diffusion features. Track4Gen merges the video genera tion and point tracking tasks into a single network by making minimal changes to existing video generation architectures. Using Stable Video Diffusion as a back bone, Track4Gen demonstrates that it is possible to unify video generation and p oint tracking, which are typically handled as separate tasks. Our extensive eval uations show that Track4Gen effectively reduces appearance drift, resulting in t emporally stable and visually coherent video generation. Project page: https://h yeonho99.github.io/track4gen/

****

Mask^2DiT: Dual Mask-based Diffusion Transformer for Multi-Scene Long Video Gene ration

Tianhao Qi, Jianlong Yuan, Wanquan Feng, Shancheng Fang, Jiawei Liu, SiYu Zhou, Qian He, Hongtao Xie, Yongdong Zhang; Proceedings of the Computer Vision and Pat tern Recognition Conference (CVPR), 2025, pp. 18837-18846

Sora has unveiled the immense potential of the Diffusion Transformer (DiT) archi tecture in single-scene video generation. However, the more challenging task of multi-scene video generation, which offers broader applications, remains relativ ely underexplored. To bridge this gap, we propose Mask^2DiT, a novel approach th at establishes fine-grained, one-to-one alignment between video segments and the ir corresponding text annotations. Specifically, we introduce a symmetric binary mask at each attention layer within the DiT architecture, ensuring that each te xt annotation applies exclusively to its respective video segment while preservi ng temporal coherence across visual tokens. This attention mechanism enables pre cise segment-level textual-to-visual alignment, allowing the DiT architecture to

effectively handle video generation tasks with a fixed number of scenes. To further equip the DiT architecture with the ability to generate additional scenes based on existing ones, we incorporate a segment-level conditional mask, which conditions each newly generated segment on the preceding video segments, thereby enabling auto-regressive scene extension. Both qualitative and quantitative experiments confirm that Mask^2DiT excels in maintaining visual consistency across segments while ensuring semantic alignment between each segment and its corresponding text description. Our project page is \href https://tianhao-qi.github.io/Mask2DiTProject/ https://tianhao-qi.github.io/Mask2DiTProject/ .
****

JamMa: Ultra-lightweight Local Feature Matching with Joint Mamba
Xiaoyong Lu, Songlin Du; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14934-14943
Existing state-of-the-art feature matchers capture long-range dependencies with Transformers but are hindered by high spatial complexity,leading to demanding training and high-latency inference.Striking a better balance between performance and efficiency remains a critical challenge in feature matching.Inspired by the linear complexity \mathcal O (N) of Mamba, we propose an ultra-lightweight Mamba-based matcher, named JamMa, which converges on a single GPU and achieves an impressive performance-efficiency balance in inference.To unlock the potential of Mamba for feature matching,we propose Joint Mamba with a scan-merge strategy named JEGO, which enables:(1) Joint scan of two images to achieve high-frequency mutual interaction, (2) Efficient scan with skip steps to reduce sequence length, (3) Global receptive field, and (4) Omnidirectional feature representation.With the above properties, the JEGO strategy significantly outperforms the scan-merge strategies proposed in VMamba and EVMamba in the feature matching task.Compared to attention-based sparse and semi-dense matchers, JamMa demonstrates a notably superior balance between performance and efficiency,delivering better performance with less than 50% of the parameters and FLOPs.
****

DyCoke: Dynamic Compression of Tokens for Fast Video Large Language Models
Keda Tao, Can Qin, Haoxuan You, Yang Sui, Huan Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18992-19001
Video large language models (VLLMs) have significantly advanced recently in processing complex video content. Yet, their inference efficiency remains constrained because of the high computational cost stemming from the thousands of visual tokens generated from the video inputs. We empirically observe that, unlike single image inputs, VLLMs typically attend visual tokens from different frames at different decoding iterations. This makes a one-shot pruning strategy prone to removing important tokens by mistake. Motivated by this, we present DyCoke, a training-free token compression method to optimize token representation and accelerate VLLMs. DyCoke incorporates a plug-and-play temporal compression module to minimize temporal redundancy by merging redundant tokens across frames and applying dynamic KV cache reduction to prune spatially redundant tokens selectively. It ensures high-quality inference by dynamically retaining the critical tokens at each decoding step. Extensive experimental results demonstrate that DyCoke can outperform the prior SoTA counterparts, achieving 1.5x inference speedup, and 1.4x memory reduction against the baseline VLLM, while still improving the performance, with no training.
****

T-FAKE: Synthesizing Thermal Images for Facial Landmarking
Philipp Flotho, Moritz Piening, Anna Kukleva, Gabriele Steidl; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26356-26366
Facial analysis is a key component in a wide range of applications such as security, autonomous driving, entertainment, and healthcare. Despite the availability of various facial RGB datasets, the thermal modality, which plays a crucial role in life sciences, medicine, and biometrics, has been largely overlooked. To address this gap, we introduce the T-FAKE dataset, a new large-scale synthetic thermal dataset with sparse and dense landmarks. To facilitate the creation of the

dataset, we propose a novel RGB2Thermal loss function, which enables the domain-adaptive transfer of thermal style to RGB faces. By utilizing the Wasserstein distance between thermal and RGB patches and the statistical analysis of clinical temperature distributions on faces, we ensure that the generated thermal images closely resemble real samples. Using RGB2Thermal style transfer based on our RGB2Thermal loss function, we create the large-scale synthetic thermal T-FAKE dataset. Leveraging our novel T-FAKE dataset, probabilistic landmark prediction, and label adaptation networks, we demonstrate significant improvements in landmark detection methods on thermal images across different landmark conventions. Our models show excellent performance with both sparse 70-point landmarks and dense 478-point landmark annotations

****

Multi-Resolution Pathology-Language Pre-training Model with Text-Guided Visual Representation

Shahad Albastaki, Anabia Sohail, Iyyakutti Iyappan Ganapathi, Basit Alawode, Asim Khan, Sajid Javed, Naoufel Werghi, Mohammed Bennamoun, Arif Mahmood; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25907-25919

In Computational Pathology (CPath), the introduction of Vision-Language Models (VLMs) has opened new avenues for research, focusing primarily on aligning image-text pairs at a single magnification level. However, this approach might not be sufficient for tasks like cancer subtype classification, tissue phenotyping, and survival analysis due to the limited level of detail that a single-resolution image can provide. Addressing this, we propose a novel multi-resolution paradigm leveraging Whole Slide Images (WSIs) to extract histology patches at multiple resolutions and generate corresponding textual descriptions through advanced CPath VLM. We introduce visual-textual alignment at multiple resolutions as well as cross-resolution alignment to establish more effective text-guided visual representations. Cross-resolution alignment using a multi-modal encoder enhances the model's ability to capture context from multiple resolutions in histology images. Our model aims to capture a broader range of information, supported by novel loss functions, enriches feature representation, improves discriminative ability, and enhances generalization across different resolutions. Pre-trained on a comprehensive TCGA dataset with 34 million image-language pairs at various resolutions, our fine-tuned model outperforms State-Of-The-Art (SOTA) counterparts across multiple datasets and tasks, demonstrating its effectiveness in CPath. The code is available on GitHub at: https://github.com/BasitAlawode/MR-PLIP.

****

InteractAnything: Zero-shot Human Object Interaction Synthesis via LLM Feedback and Object Affordance Parsing

Jinlu Zhang, Yixin Chen, Zan Wang, Jie Yang, Yizhou Wang, Siyuan Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7015-7025

Recent advances in 3D human-aware generation have made significant progress. However, existing methods still struggle with generating novel Human Object Interaction (HOI) from text, particularly for open-set objects. We identify three main challenges of this task: precise human-object relation reasoning, affordance parsing for any object, and detailed human interaction pose synthesis aligning description and object geometry. In this work, we propose a novel zero-shot 3D HOI generation framework without training on specific datasets, leveraging the knowledge from large-scale pre-trained models. Specifically, the human-object relations are inferred from large language models (LLMs) to initialize object properties and guide the optimization process. Then we utilize a pre-trained 2D image diffusion model to parse unseen objects and extract contact points, avoiding the limitations imposed by existing 3D asset knowledge. The initial human pose is generated by sampling multiple hypotheses through multi-view SDS based on the input text and object geometry. Finally, we introduce a detailed optimization to generate fine-grained, precise, and natural interaction, enforcing realistic 3D contact between the 3D object and the involved body parts, including hands in grasping. This is achieved by distilling human-level feedback from LLMs to capture detai

led human-object relations from the text instruction. Extensive experiments validate the effectiveness of our approach compared to prior works, particularly in terms of the fine-grained nature of interactions and the ability to handle open-set 3D objects. Project page: jinluzhang.site/projects/interactanything.
****

MammAlps: A Multi-view Video Behavior Monitoring Dataset of Wild Mammals in the Swiss Alps
Valentin Gabeff, Haozhe Qi, Brendan Flaherty, Gencer Sumbul, Alexander Mathis, Devis Tuia; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13854-13864
Monitoring wildlife is essential for ecology and ethology, especially in light of the increasing human impact on ecosystems. Camera traps have emerged as habitat-centric sensors enabling the study of wildlife populations at scale with minimal disturbance. However, the lack of annotated video datasets limits the development of powerful video understanding models needed to process the vast amount of fieldwork data collected. To advance research in wild animal behavior monitoring we present MammAlps, a multimodal and multi-view dataset of wildlife behavior monitoring from 9 camera-traps in the Swiss National Park. MammAlps contains over 14 hours of video with audio, 2D segmentation maps and 8.5 hours of individual tracks densely labeled for species and behavior. Based on 6135 single animal clips, we propose the first hierarchical and multimodal animal behavior recognition benchmark using audio, video and reference scene segmentation maps as inputs. Furthermore, we also propose a second ecology-oriented benchmark aiming at identifying activities, species, number of individuals and meteorological conditions from 397 multi-view and long-term ecological events, including false positive triggers. We advocate that both tasks are complementary and contribute to bridging the gap between machine learning and ecology. Code and data are available at: https://github.com/eceo-epfl/MammAlps
****

Diffusion-based Realistic Listening Head Generation via Hybrid Motion Modeling
Yinuo Wang, Yanbo Fan, Xuan Wang, Guo Yu, Fei Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15885-15895
Listening head generation aims to synthesize non-verbal responsive listening head videos that naturally react to a certain speaker, for which, both realistic head movements, expressive facial expressions, and high visual qualities are expected. Previous approaches typically follow a two-stage pipeline that first generates intermediate 3D motion signals such as 3DMM coefficients, and then synthesizes the videos by deterministic rendering, suffering from limited motion expressiveness and low visual quality (eg, 256 x256). In this work, we propose a novel listening head generation method that harnesses the generative capabilities of the diffusion model for both motion generation and high-quality rendering. Crucially, we propose an effective hybrid motion modeling module that addresses training difficulties caused by the scarcity of listening head data while preserving the intricate details that may be lost in explicit motion representations. We further develop a tailored control guidance for head pose and facial expression, by integrating their intrinsic motion characteristics. Our method enables high-fidelity video generation with 512x512 resolution and delivers vivid listener motion feedback. We conduct comprehensive experiments and obtain superior performance in terms of both visual quality and motion expressiveness compared with existing methods.
****

SAT-HMR: Real-Time Multi-Person 3D Mesh Estimation via Scale-Adaptive Tokens
Chi Su, Xiaoxuan Ma, Jiajun Su, Yizhou Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16796-16806
We propose a one-stage framework for real-time multi-person 3D human mesh estimation from a single RGB image. While current one-stage methods, which follow a DETR-style pipeline, achieve state-of-the-art (SOTA) performance with high-resolution inputs, we observe that this particularly benefits the estimation of individuals in smaller scales of the image (e.g., those far from the camera), but at the cost of significantly increased computation overhead. To address this, we intr

oduce scale-adaptive tokens that are dynamically adjusted based on the relative scale of each individual in the image within the DETR framework. Specifically, individuals in smaller scales are processed at higher resolutions, larger ones at lower resolutions, and background regions are further distilled. These scale-adaptive tokens more efficiently encode the image features, facilitating subsequent decoding to regress the human mesh, while allowing the model to allocate computational resources more effectively and focus on more challenging cases. Experiments show that our method preserves the accuracy benefits of high-resolution processing while substantially reducing computational cost, achieving real-time inference with performance comparable to SOTA methods. Code and models are available at https://ChiSu001.github.io/SAT-HMR/.

****

PICD: Versatile Perceptual Image Compression with Diffusion Rendering

Tongda Xu, Jiahao Li, Bin Li, Yan Wang, Ya-Qin Zhang, Yan Lu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28436-28445

Recently, perceptual image compression has achieved significant advancements, delivering high visual quality at low bitrates for natural images. However, for screen content, existing methods often produce noticeable artifacts when compressing text. To tackle this challenge, we propose versatile perceptual screen image compression with diffusion rendering (PICD), a codec that works well for both screen and natural images. More specifically, we propose a compression framework that encodes the text and image separately, and renders them into one image using diffusion model. For this diffusion rendering, we integrate conditional information into diffusion models at three distinct levels: 1). Domain level: We fine-tune the base diffusion model using text content prompts with screen content. 2). Adaptor level: We develop an efficient adaptor to control the diffusion model using compressed image and text as input. 3). Instance level: We apply instance-wise guidance to further enhance the decoding process. Empirically, our PICD surpasses existing perceptual codecs in terms of both text accuracy and perceptual quality. Additionally, without text conditions, our approach serves effectively as a perceptual codec for natural images.

****

UniScene: Unified Occupancy-centric Driving Scene Generation

Bohan Li, Jiazhe Guo, Hongsi Liu, Yingshuang Zou, Yikang Ding, Xiwu Chen, Hu Zhu, Feiyang Tan, Chi Zhang, Tiancai Wang, Shuchang Zhou, Li Zhang, Xiaojuan Qi, Hao Zhao, Mu Yang, Wenjun Zeng, Xin Jin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11971-11981

Generating high-fidelity, controllable, and annotated training data is critical for autonomous driving. Existing methods typically generate a single data form directly from a coarse scene layout, which not only fails to output rich data forms required for diverse downstream tasks but also struggles to model the direct layout-to-data distribution. In this paper, we introduce UniScene, the first unified framework for generating three key data forms - semantic occupancy, video, and LiDAR - in driving scenes. UniScene employs a progressive generation process that decomposes the complex task of scene generation into two hierarchical steps: (a) first generating semantic occupancy from a customized scene layout as a meta scene representation rich in both semantic and geometric information, and then (b) conditioned on occupancy, generating video and LiDAR data, respectively, with two novel transfer strategies of Gaussian-based Joint Rendering and Prior-guided Sparse Modeling. This occupancy-centric approach reduces the generation burden, especially for intricate scenes, while providing detailed intermediate representations for the subsequent generation stages. Extensive experiments demonstrate that UniScene outperforms previous SOTAs in the occupancy, video, and LiDAR generation, which also indeed benefits downstream driving tasks. The Project is available at https://arlo0o.github.io/uniscene/.

****

Wonderland: Navigating 3D Scenes from a Single Image

Hanwen Liang, Junli Cao, Vidit Goel, Guocheng Qian, Sergei Korolev, Demetri Terzopoulos, Konstantinos N. Plataniotis, Sergey Tulyakov, Jian Ren; Proceedings of

the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 798-810
This paper addresses a challenging question: how can we efficiently create high-quality, wide-scope 3D scenes from a single arbitrary image?Existing methods face several constraints, such as requiring multi-view data, time-consuming per-scene optimization, low visual quality, and distorted reconstructions for unseen areas. We propose a novel pipeline to overcome these limitations.Specifically, we introduce a large-scale reconstruction model that uses latents from a video diffusion model to predict 3D Gaussian Splatting, even from a single-condition image, in a feed-forward manner. The video diffusion model is designed to precisely follow a specified camera trajectory, allowing it to generate compressed latents that contain multi-view information while maintaining 3D consistency.We further train the 3D reconstruction model to operate on the video latent space with a progressive training strategy, enabling the generation of high-quality, wide-scope, and generic 3D scenes.Extensive evaluations on various datasets show that our model significantly outperforms existing methods for single-view 3D rendering, particularly with out-of-domain images. For the first time, we demonstrate that a 3D reconstruction model can be effectively built upon the latent space of a diffusion model to realize efficient 3D scene generation.
****

Learning from Streaming Video with Orthogonal Gradients
Tengda Han, Dilara Gokay, Joseph Heyward, Chuhan Zhang, Daniel Zoran, Viorica Patraucean, Joao Carreira, Dima Damen, Andrew Zisserman; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13651-13660
We address the challenge of representation learning from a continuous stream of video as input, in a self-supervised manner. This differs from the standard approaches to video learning where videos are chopped and shuffled during training in order to create a non-redundant batch that satisfies the independently and identically distributed (IID) sample assumption expected by conventional training paradigms. When videos are only available as a continuous stream of input, the IID assumption is evidently broken, leading to poor performance. We demonstrate the drop in performance when moving from shuffled to sequential learning on three tasks: the one-video representation learning method DoRA, standard VideoMAE on multi-video datasets, and the task of future video prediction. To address this drop, we propose a geometric modification to standard optimizers, to decorrelate batches by utilising orthogonal gradients during training. The proposed modification can be applied to any optimizer -- we demonstrate it with Stochastic Gradient Descent (SGD) and AdamW. Our proposed orthogonal optimizer allows models trained from streaming videos to alleviate the drop in representation learning performance, as evaluated on downstream tasks. On three scenarios (DoRA, VideoMAE, future prediction), we show our orthogonal optimizer outperforms the strong AdamW in all three scenarios.
****

Towards Satellite Image Road Graph Extraction: A Global-Scale Dataset and A Novel Method
Pan Yin, Kaiyu Li, Xiangyong Cao, Jing Yao, Lei Liu, Xueru Bai, Feng Zhou, Deyu Meng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1527-1537
Recently, road graph extraction has garnered increasing attention due to its crucial role in autonomous driving, navigation, etc. However, accurately and efficiently extracting road graphs remains a persistent challenge, primarily due to the severe scarcity of labeled data. To address this limitation, we collect a global-scale satellite road graph extraction dataset, i.e. Global-Scale dataset. Specifically, the Global-Scale dataset is 20x larger than the largest existing public road extraction dataset and spans over 13,800 km^2 globally. Additionally, we develop a novel road graph extraction model, i.e. SAM-Road++, which adopts a node-guided resampling method to alleviate the mismatch issue between training and inference in SAM-Road, a pioneering state-of-the-art road graph extraction model. Furthermore, we propose a simple yet effective "extended-line" strategy in SAM-Road++ to mitigate the occlusion issue on the road. Extensive experiments demonstrate the validity of the collected Global-Scale dataset and the proposed SAM

-Road++ method, particularly highlighting its superior predictive power in unsee
n regions.
****

SuperLightNet: Lightweight Parameter Aggregation Network for Multimodal Brain Tu
mor Segmentation

Feng Yu, Jiacheng Cao, Li Liu, Minghua Jiang; Proceedings of the Computer Vision
 and Pattern Recognition Conference (CVPR), 2025, pp. 5197-5206

Multimodal 3D segmentation involves a significant number of 3D convolution opera
tions, which requires substantial computational resources and high-performance c
omputing devices in MRI multimodal brain tumor segmentation. The key challenge i
n multimodal 3D segmentation is how to minimize network computational load while
 maintaining high accuracy. To address the issue, a novel lightweight parameter
aggregation network (SuperLightNet) is proposed to realize the efficient encoder
 and decoder for the high accurate and low computation. A random multiview drop
encoder is designed to learn the spatial structure of multimodal images through
a random multi-view approach for solving the high computational time complexity
that has arisen in recent years with methods relying on transformers and Mamba.
A learnable residual skip decoder is designed to incorporate learnable residual
and group skip weights for addressing the reduced computational efficiency cause
d by the use of overly heavy convolution and deconvolution decoders. Experimenta
l results demonstrate that the proposed method achieves a leading reduction in p
arameter count by 95.59%, the 96.78% improvement in computational efficiency, th
e 96.86% enhancement in memory access performance, and the average performance g
ain of 0.21% on the BraTS2019 and BraTS2021 datasets in comparison with the stat
e-of-the-art methods. Code is available at https://github.com/WTU1020-Medical-Se
gmentation/SuperLightNet.
****

VideoSPatS: Video SPatiotemporal Splines for Disentangled Occlusion, Appearance
 and Motion Modeling and Editing

Juan Luis Gonzalez, Xu Yao, Alex Whelan, Kyle Olszewski, Hyeongwoo Kim, Pablo Ga
rrido; Proceedings of the Computer Vision and Pattern Recognition Conference (CV
PR), 2025, pp. 22901-22910

We present an implicit video representation for occlusions, appearance, and moti
on disentanglement from monocular videos, which we refer to as Video Spatiotempo
ral Splines (VideoSPatS).Unlike previous methods that map time and coordinates t
o deformation and canonical colors, our VideoSPatS maps input coordinates into S
patial and Color Spline deformation fields Ds and Dc, which disentangle motion a
nd appearance in videos. With spline-based parametrization, our method naturally
 generates temporally consistent flow and guarantees long-term temporal consiste
ncy, which is crucial for convincing video editing.Aided by additional predictio
n blocks, our VideoSPatS also performs layer separation between the latent video
 and the selected occluder. By disentangling occlusions, appearance, and motion,
 our method allows for better spatiotemporal modeling and editing of diverse vid
eos, including in-the-wild talking head videos with challenging occlusions, shad
ows, and specularities while maintaining a reasonable canonical space for editin
g.We also present general video modeling results on the DAVIS, and CoDeF dataset
s, as well as our own talking head video dataset collected from open-source web
videos. Extensive ablations show the combination of Ds and Dc under neural spli
nes can overcome motion and appearance ambiguities, paving the way to more advan
ced video editing models.
****

Classifier-to-Bias: Toward Unsupervised Automatic Bias Detection for Visual Clas
sifiers

Quentin Guimard, Moreno D'Incà, Massimiliano Mancini, Elisa Ricci; Proceedings o
f the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15151
-15161

A person downloading a pre-trained model from the web should be aware of its bia
ses. Existing approaches for bias identification rely on datasets containing lab
els for the task of interest, something that a non-expert may not have access to
, or may not have the necessary resources to collect: this greatly limits the nu

mber of tasks where model biases can be identified. In this work, we present Cla ssifier-to-Bias (C2B), the first bias discovery framework that works without acc ess to any labeled data: it only relies on a textual description of the classifi cation task to identify biases in the target classification model. This descript ion is fed to a large language model to generate bias proposals and correspondin g captions depicting biases together with task-specific target labels. A retriev al model collects images for those captions, which are then used to assess the a ccuracy of the model w.r.t. the given biases. C2B is training-free, does not req uire any annotations, has no constraints on the list of biases, and can be appli ed to any pre-trained model on any classification task. Experiments on two publi cly available datasets show that C2B discovers biases beyond those of the origin al datasets and outperforms a recent state-of-the-art bias detection baseline th at relies on task-specific annotations, being a promising first step toward addr essing task-agnostic unsupervised bias detection.

****

Self-Supervised Spatial Correspondence Across Modalities

Ayush Shrivastava, Andrew Owens; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6383-6393

We present a method for finding cross-modal space-time correspondences. Given t wo images from different visual modalities, such as an RGB image and a depth map , our model identifies which pairs of pixels correspond to the same physical poi nts in the scene. To solve this problem, we extend the contrastive random walk f ramework to simultaneously learn cycle-consistent feature representations for bo th cross-modal and intra-modal matching. The resulting model is simple and has n o explicit photo-consistency assumptions. It can be trained entirely using unlab eled data, without the need for any spatially aligned multimodal image pairs. We evaluate our method on both geometric and semantic correspondence tasks. For ge ometric matching, we consider challenging tasks such as RGB-to-depth and RGB-to- thermal matching (and vice versa); for semantic matching, we evaluate on photo-s ketch and cross-style image alignment. Our method achieves strong performance ac ross all benchmarks.

****

MOS-Attack: A Scalable Multi-objective Adversarial Attack Framework

Ping Guo, Cheng Gong, Xi Lin, Fei Liu, Zhichao Lu, Qingfu Zhang, Zhenkun Wang; P roceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 202 5, pp. 5041-5051

Crafting adversarial examples is crucial for evaluating and enhancing the robust ness of Deep Neural Networks (DNNs), presenting a challenge equivalent to maximi zing a non-differentiable 0-1 loss function. However, existing single objecti ve methods, namely adversarial attacks focus on a surrogate loss function, do no t fully harness the benefits of engaging multiple loss functions, as a result of insufficient understanding of their synergistic and conflicting nature. To o vercome these limitations, we propose the Multi-Objective Set-based Attack (MOS Attack), a novel adversarial attack framework leveraging multiple loss functions and automatically uncovering their interrelations. The MOS Attack adopts a s et-based multi-objective optimization strategy, enabling the incorporation of nu merous loss functions without additional parameters. It also automatically mi nes synergistic patterns among various losses, facilitating the generation of po tent adversarial attacks with fewer objectives. Extensive experiments have sh own that our MOS Attack outperforms single-objective attacks. Furthermore, by ha rnessing the identified synergistic patterns, MOS Attack continues to show super ior results with a reduced number of loss functions.

****

Six-CD: Benchmarking Concept Removals for Text-to-image Diffusion Models

Jie Ren, Kangrui Chen, Yingqian Cui, Shenglai Zeng, Hui Liu, Yue Xing, Jiliang T ang, Lingjuan Lyu; Proceedings of the Computer Vision and Pattern Recognition Co nference (CVPR), 2025, pp. 28769-28778

Text-to-image (T2I) diffusion models have shown exceptional capabilities in gene rating images that closely correspond to textual prompts. However, the advanceme nt of T2I diffusion models presents significant risks, as the models could be ex

ploited for malicious purposes, such as generating images with violence or nudity, or creating unauthorized portraits of public figures in inappropriate contexts. To mitigate these risks, concept removal methods have been proposed. These methods aim to modify diffusion models to prevent the generation of malicious and unwanted concepts. Despite these efforts, existing research faces several challenges: (1) a lack of consistent comparisons on a comprehensive dataset, (2) ineffective prompts in harmful and nudity concepts, (3) overlooked evaluation of the ability to generate the benign part within prompts containing malicious concepts. To address these gaps, we propose to benchmark the concept removal methods by introducing a new dataset, Six-CD, along with a novel evaluation metric. In this benchmark, we conduct a thorough evaluation of concept removals, with the experimental observations and discussions offering valuable insights in the field.

****

Motion Modes: What Could Happen Next?

Karran Pandey, Yannick Hold-Geoffroy, Matheus Gadelha, Niloy J. Mitra, Karan Singh, Paul Guerrero; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2030-2039

Predicting diverse object motions from a single static image remains challenging, as current video generation models often entangle object movement with camera motion and other scene changes. While recent methods can predict specific motions from motion arrow input, they rely on synthetic data and predefined motions, limiting their application to complex scenes. We introduce Motion Modes, a training-free approach that explores a pre-trained image-to-video generator's latent distribution to discover various distinct and plausible motions focused on selected objects in static images. We achieve this by employing a flow generator guided by energy functions designed to disentangle object and camera motion. Additionally, we use an energy inspired by particle guidance to diversify the generated motions, without requiring explicit training data. Experimental results demonstrate that Motion Modes generates realistic and varied object animations, surpassing previous methods and even human predictions regarding plausibility and diversity.

****

Finer-CAM: Spotting the Difference Reveals Finer Details for Visual Explanation

Ziheng Zhang, Jianyang Gu, Arpita Chowdhury, Zheda Mai, David Carlyn, Tanya Berger-Wolf, Yu Su, Wei-Lun Chao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9611-9620

Class activation map (CAM) has been widely used to highlight image regions that contribute to class predictions. Despite its simplicity and computational efficiency, CAM often struggles to identify discriminative regions that distinguish visually similar fine-grained classes. Prior efforts address this limitation by introducing more sophisticated explanation processes, but at the cost of extra complexity. In this paper, we propose Finer-CAM, a method that retains CAM's efficiency while achieving precise localization of discriminative regions. Our key insight is that the deficiency of CAM lies not in "how" it explains, but in "what" it explains. Specifically, previous methods attempt to identify all cues contributing to the target class's logit value, which inadvertently also activates regions predictive of visually similar classes. By explicitly comparing the target class with similar classes and spotting their differences, Finer-CAM suppresses features shared with other classes and emphasizes the unique, discriminative details of the target class. Finer-CAM is easy to implement, compatible with various CAM methods, and can be extended to multi-modal models for accurate localization of specific concepts. Additionally, Finer-CAM allows adjustable comparison strength, enabling users to selectively highlight coarse object contours or fine discriminative details. Quantitatively, we show that masking out the top 5% of activated pixels by Finer-CAM results in a larger relative confidence drop compared to baselines. The source code and demo are available at https://github.com/Imageomics/Finer-CAM.

****

The Change You Want To Detect: Semantic Change Detection In Earth Observation With Hybrid Data Generationf

Yanis Benidir, Nicolas Gonthier, Clement Mallet; Proceedings of the Computer Vis
ion and Pattern Recognition Conference (CVPR), 2025, pp. 2204-2214
Bi-temporal change detection at scale based on Very High Resolution (VHR) images
 is crucial for Earth monitoring. Such task remains poorly addressed even in the
 deep learning era: it either requires large volumes of annotated data - in the
semantic case - or is limited to restricted datasets for binary set-ups. Most ap
proaches do not exhibit the versatility required for temporal and spatial adapta
tion: simplicity in architecture design and pretraining on realistic and compreh
ensive datasets. Synthetic datasets is the key solution but still fails handling
 complex and diverse scenes. In this paper, we present HySCDG a generative pipel
ine for creating a large hybrid semantic change detection dataset that contains
both real VHR images and inpainted ones, along with land cover semantic map at b
oth dates and the change map. Being semantically and spatially guided, HySCDG ge
nerates realistic images, leading to a comprehensive and hybrid transfer-proof d
ataset FSC-180k. We evaluate FSC-180k on five change detection cases (binary and
 semantic), from zero-shot to mixed and sequential training, and also under low
data regime training. Experiments demonstrate that pretraining on our hybrid dat
aset leads to a significant performance boost, outperforming SyntheWorld, a full
y synthetic dataset, in every configuration. All codes, models, and data will be
 made available.
****

Black-Box Forgery Attacks on Semantic Watermarks for Diffusion Models
Andreas Müller, Denis Lukovnikov, Jonas Thietke, Asja Fischer, Erwin Quiring; Pr
oceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025
, pp. 20937-20946
Integrating watermarking into the generation process of latent diffusion models
(LDMs) simplifies detection and attribution of generated content. Semantic water
marks, such as Tree-Rings and Gaussian Shading, represent a novel class of water
marking techniques that are easy to implement and highly robust against various
perturbations. However, our work demonstrates a fundamental security vulnerabili
ty of semantic watermarks. We show that attackers can leverage unrelated models,
 even with different latent spaces and architectures (UNet vs DiT), to perform p
owerful and realistic forgery attacks. Specifically, we design two watermark for
gery attacks. The first \imprints a targeted watermark into real images by manip
ulating the latent representation of an arbitrary image in an unrelated LDM to g
et closer to the latent representation of a watermarked image. We also show that
 this technique can be used for watermark removal. The second attack generates n
ew images with the target watermark by inverting a watermarked image and re-gene
rating it with an arbitrary prompt. Both attacks just need a single reference im
age with the target watermark. Overall, our findings question the applicability
of semantic watermarks by revealing that attackers can easily forge or remove th
ese watermarks under realistic conditions.
****

An Image-like Diffusion Method for Human-Object Interaction Detection
Xiaofei Hui, Haoxuan Qu, Hossein Rahmani, Jun Liu; Proceedings of the Computer V
ision and Pattern Recognition Conference (CVPR), 2025, pp. 14002-14012
Human-object interaction (HOI) detection often faces high levels of ambiguity an
d indeterminacy, as the same interaction can appear vastly different across diff
erent human-object pairs. Additionally, the indeterminacy can be further exacerb
ated by issues such as occlusions and cluttered backgrounds. To handle such a ch
allenging task, in this work, we begin with a key observation: the output of HOI
 detection for each human-object pair can be recast as an image. Thus, inspired
by the strong image generation capabilities of image diffusion models, we propos
e a new framework, HOI-IDiff. In HOI-IDiff, we tackle HOI detection from a novel
 perspective, using an Image-like Diffusion process to generate HOI detection ou
tputs as images. Furthermore, recognizing that our recast images differ in certa
in properties from natural images, we enhance our framework with a customized HO
I diffusion process and a slice patchification model architecture, which are spe
cifically tailored to generate our recast "HOI images". Extensive experiments de
monstrate the efficacy of our framework.

****
VidSeg: Training-free Video Semantic Segmentation based on Diffusion Models
Qian Wang, Abdelrahman Eldesokey, Mohit Mendiratta, Fangneng Zhan, Adam Kortylewski, Christian Theobalt, Peter Wonka; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22985-22994
We introduce the first training-free approach for Video Semantic Segmentation (VSS) based on pre-trained diffusion models. A growing research direction attempts to employ diffusion models to perform downstream vision tasks by exploiting their deep understanding of image semantics. Yet, the majority of these approaches have focused on image-related tasks like semantic segmentation, with less emphasis on video tasks such as VSS. Ideally, diffusion-based image semantic segmentation approaches can be applied to videos in a frame-by-frame manner. However, we find their performance on videos to be subpar due to the absence of any modeling of temporal information inherent in the video data. To this end, we tackle this problem and introduce a framework tailored for VSS based on pre-trained image and video diffusion models. We propose building a scene context model based on the diffusion features, where the model is autoregressively updated to adapt to scene changes. This context model predicts per-frame coarse segmentation maps that are temporally consistent. To refine these maps further, we propose a correspondence-based refinement strategy that aggregates predictions temporally, resulting in more confident predictions. Finally, we introduce a masked modulation approach to upsample the coarse maps to a high-quality full resolution. Experiments show that our proposed approach significantly outperforms existing training-free image semantic segmentation approaches on various VSS benchmarks without any training or fine-tuning. Moreover, it rivals supervised VSS approaches on the VSPW dataset despite not being explicitly trained for VSS.
****
COB-GS: Clear Object Boundaries in 3DGS Segmentation Based on Boundary-Adaptive Gaussian Splitting
Jiaxin Zhang, Junjun Jiang, Youyu Chen, Kui Jiang, Xianming Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19335-19344
Accurate object segmentation is crucial for high-quality scene understanding in the 3D vision domain. However, 3D segmentation based on 3D Gaussian Splatting (3DGS) struggles with accurately delineating object boundaries, as Gaussian primitives often span across object edges due to their inherent volume and the lack of semantic guidance during training. In order to tackle these challenges, we introduce Clear Object Boundaries for 3DGS Segmentation (COB-GS), which aims to improve segmentation accuracy by clearly delineating blurry boundaries of interwoven Gaussian primitives within the scene. Unlike existing approaches that remove ambiguous Gaussians and sacrifice visual quality, COB-GS, as a 3DGS refinement method, jointly optimizes semantic and visual information, allowing the two different levels to cooperate with each other effectively. Specifically, for the semantic guidance, we introduce a boundary-adaptive Gaussian splitting technique that leverages semantic gradient statistics to identify and split ambiguous Gaussians, aligning them closely with object boundaries. For the visual optimization, we rectify the degraded suboptimal texture of the 3DGS scene, particularly along the refined boundary structures. Experimental results show that COB-GS substantially improves segmentation accuracy and robustness against inaccurate masks from pre-trained model, yielding clear boundaries while preserving high visual quality. Code is available at https://github.com/ZestfulJX/COB-GS.
****
Weakly Supervised Semantic Segmentation via Progressive Confidence Region Expansion
Xiangfeng Xu, Pinyi Zhang, Wenxuan Huang, Yunhang Shen, Haosheng Chen, Jingzhong Lin, Wei Li, Gaoqi He, Jiao Xie, Shaohui Lin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9829-9838
Weakly supervised semantic segmentation (WSSS) has garnered considerable attention due to its effective reduction of annotation costs. Most approaches utilize Class Activation Maps (CAM) to produce pseudo-labels, thereby localizing target r

egions using only image-level annotations. However, the prevalent methods relyin g on vision transformers (ViT) encounter an "over-expansion" issue, i.e., CAM in correctly expands high activation value from the target object to the background regions, as it is difficult to learn pixel-level local intrinsic inductive bias in ViT from weak supervisions. To solve this problem, we propose a Progressive Confidence Region Expansion (PCRE) framework for WSSS, it gradually learns a fai thful mask over the target region and utilizes this mask to correct the confusio n in CAM. PCRE has two key components: Confidence Region Mask Expansion (CRME) a nd Class-Prototype Enhancement (CPE). CRME progressively expands the mask in the small region with the highest confidence, eventually encompassing the entire ta rget, thereby avoiding unintended CPE aims to enhance mask generation in CRME by leveraging the similarity between the learned, dataset-level class prototypes a nd patch features as supervision to optimize the mask output from CRME. Extensiv e experiments demonstrate that our method outperforms the existing single-stage and multi-stage approaches on the PASCAL VOC and MS COCO benchmark datasets. Our code is available at https://github.com/xxf011/WSSS-PCRE .

****

RELOCATE: A Simple Training-Free Baseline for Visual Query Localization Using Re gion-Based Representations

Savya Khosla, Sethuraman T V, Alexander Schwing, Derek Hoiem; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3697-3706

We present RELOCATE, a simple training-free baseline designed to perform the cha llenging task of visual query localization in long videos. To eliminate the need for task-specific training and efficiently handle long videos, RELOCATE leverag es a region-based representation derived from pretrained vision models. At a hig h level, it follows the classic object localization approach: (1) identify all o bjects in each video frame, (2) compare the objects with the given query and sel ect the most similar ones, and (3) perform bidirectional tracking to get a spati o-temporal response. However, we propose some key enhancements to handle small o bjects, cluttered scenes, partial visibility, and varying appearances. Notably, we refine the selected objects for accurate localization and generate additional visual queries to capture visual variations. We evaluate RELOCATE on the challe nging Ego4D Visual Query 2D Localization dataset, establishing a new baseline th at outperforms prior task-specific methods by 49% (relative improvement) in spat io-temporal average precision.

****

PEER Pressure: Model-to-Model Regularization for Single Source Domain Generaliza tion

Dong Kyu Cho, Inwoo Hwang, Sanghack Lee; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15360-15370

Data augmentation is a popular tool for single source domain generalization, whi ch expands the source domain by generating simulated ones, improving generalizat ion on unseen target domains. In this work, we show that the performance of such augmentation-based methods in the target domains universally fluctuates during training, posing challenges in model selection under realistic scenarios. We arg ue that the fluctuation stems from the inability of the model to accumulate the knowledge learned from diverse augmentations, exacerbating feature distortion du ring training. Based on this observation, we propose a novel generalization meth od, coined Parameter-Space Ensemble with Entropy Regularization (PEER), that use s a proxy model to learn the augmented data on behalf of the main model. The mai n model is updated by averaging its parameters with the proxy model, progressive ly accumulating knowledge over the training steps. Maximizing the mutual informa tion between the output representations of the two models guides the learning pr ocess of the proxy model, mitigating feature distortion during training. Experim ental results demonstrate the effectiveness of PEER in reducing the OOD performa nce fluctuation and enhancing generalization across various datasets, including PACS, Digits, Office-Home, and VLCS. Notably, our method with simple random augm entation achieves state-of-the-art performance, surpassing prior approaches on s DG that utilize complex data augmentation strategies.

****

HOTFormerLoc: Hierarchical Octree Transformer for Versatile Lidar Place Recognition Across Ground and Aerial Views

Ethan Griffiths, Maryam Haghighat, Simon Denman, Clinton Fookes, Milad Ramezani; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6648-6658

We present HOTFormerLoc, a novel and versatile Hierarchical Octree-based TransFormer, for large-scale 3D place recognition in both ground-to-ground and ground-to-aerial scenarios across urban and forest environments. We propose an octree-based multi-scale attention mechanism that captures spatial and semantic features across granularities. To address the variable density of point distributions from spinning lidar, we present cylindrical octree attention windows to reflect the underlying distribution during attention.We introduce relay tokens to enable efficient global-local interactions and multi-scale representation learning at reduced computational cost. Our pyramid attentional pooling then synthesises a robust global descriptor for end-to-end place recognition in challenging environments. In addition, we introduce CS-Wild-Places, a novel 3D cross-source dataset featuring point cloud data from aerial and ground lidar scans captured in dense forests. Point clouds in CS-Wild-Places contain representational gaps and distinctive attributes such as varying point densities and noise patterns, making it a challenging benchmark for cross-view localisation in the wild. HOTFormerLoc achieves a top-1 average recall improvement of 5.5% - 11.5% on the CS-Wild-Places benchmark. Furthermore, it consistently outperforms SOTA 3D place recognition methods, with an average performance gain of 4.9% on well-established urban and forest datasets. The code and CS-Wild-Places benchmark is available at https://csiro-robotics.github.io/HOTFormerLoc.

****

Revisiting Fairness in Multitask Learning: A Performance-Driven Approach for Variance Reduction

Xiaohan Qin, Xiaoxing Wang, Junchi Yan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20492-20501

Multi-task learning (MTL) can leverage shared knowledge across tasks to improve data efficiency and generalization performance, and has been applied in various scenarios. However, task imbalance remains a major challenge for existing MTL methods. While the prior works have attempted to mitigate inter-task unfairness through loss-based and gradient-based strategies, they still exhibit imbalanced performance across tasks on common benchmarks.This key observation motivates us to consider performance-level information as an explicit fairness indicator, which can precisely reflect the current optimization status of each task, and accordingly help to adjust the gradient aggregation process.Specifically, we utilize the performance variance among tasks as the fairness indicator and introduce a dynamic weighting strategy to gradually reduce the performance variance. Based on this, we propose PIVRG, a novel performance-informed variance reduction gradient aggregation approach.Extensive experiments show that PIVRG achieves SOTA performance across various benchmarks, spanning both supervised learning and reinforcement learning tasks with task numbers ranging from 2 to 40. Results from the ablation study also show that our approach can be integrated into existing methods, significantly enhancing their performance while reducing the performance variance among tasks, thus achieving fairer optimization.

****

UniK3D: Universal Camera Monocular 3D Estimation

Luigi Piccinelli, Christos Sakaridis, Mattia Segu, Yung-Hsu Yang, Siyuan Li, Wim Abbeloos, Luc Van Gool; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1028-1039

Monocular 3D estimation is crucial for visual perception. However, current methods fall short by relying on oversimplified assumptions, such as pinhole camera models or rectified images. These limitations severely restrict their general applicability, causing poor performance in real-world scenarios with fisheye or panoramic images and resulting in substantial context loss. To address this, we present UniK3D, the first generalizable method for monocular 3D estimation able to model any camera. Our method introduces a spherical 3D representation which allo

ws for better disentanglement of camera and scene geometry and enables accurate metric 3D reconstruction for unconstrained camera models. Our camera component features a novel, model-independent representation of the pencil of rays, achieved through a learned superposition of spherical harmonics. We also introduce an angular loss, which, together with the camera module design, prevents the contraction of the 3D outputs for wide-view cameras. A comprehensive zero-shot evaluation on 13 diverse datasets demonstrates the state-of-the-art performance of UniK3D across 3D, depth, and camera metrics, with substantial gains in challenging large-field-of-view and panoramic settings, while maintaining top accuracy in conventional pinhole small-field-of-view domains. Code and models are available at github.com/lpiccinelli-eth/unik3d

****

ConMo: Controllable Motion Disentanglement and Recomposition for Zero-Shot Motion Transfer

Jiayi Gao, Zijin Yin, Changcheng Hua, Yuxin Peng, Kongming Liang, Zhanyu Ma, Jun Guo, Yang Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7191-7200

The development of Text-to-Video (T2V) generation has made motion transfer possible, enabling the control of video motion based on existing footage. However, current methods have two limitations: 1) struggle to handle multi-subjects videos, failing to transfer specific subject motion; 2) struggle to preserve the diversity and accuracy of motion as transferring to subjects with varying shapes. To overcome these, we introduce ConMo, a zero-shot framework that disentangle and recompose the motions of subjects and camera movements. ConMo isolates individual subject and background motion cues from complex trajectories in source videos using only subject masks, and reassembles them for target video generation. This approach enables more accurate motion control across diverse subjects and improves performance in multi-subject scenarios. Additionally, we propose soft guidance in the recomposition stage which controls the retention of original motion to adjust shape constraints, aiding subject shape adaptation and semantic transformation. Unlike previous methods, ConMo unlocks a wide range of applications, including subject size and position editing, subject removal, semantic modifications, and camera motion simulation. Extensive experiments demonstrate that ConMo significantly outperforms state-of-the-art methods in motion fidelity and semantic consistency. The code is available at https://github.com/Andyplus1/ConMo.

****

VideoMage: Multi-Subject and Motion Customization of Text-to-Video Diffusion Models

Chi-Pin Huang, Yen-Siang Wu, Hung-Kai Chung, Kai-Po Chang, Fu-En Yang, Yu-Chiang Frank Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17603-17612

Customized text-to-video generation aims to produce high-quality videos that incorporate user-specified subject identities or motion patterns. However, existing methods mainly focus on personalizing a single concept, either subject identity or motion pattern, limiting their effectiveness for multiple subjects with the desired motion patterns. To tackle this challenge, we propose a unified framework VideoMage for video customization over both multiple subjects and their interactive motions. VideoMage employs subject and motion LoRAs to capture personalized content from user-provided images and videos, along with an appearance-agnostic motion learning approach to disentangle motion patterns from visual appearance. Furthermore, we develop a spatial-temporal composition scheme to guide interactions among subjects within the desired motion patterns. Extensive experiments demonstrate that VideoMage outperforms existing methods, generating coherent, user-controlled videos with consistent subject identities and interactions.

****

AG-VPReID: A Challenging Large-Scale Benchmark for Aerial-Ground Video-based Person Re-Identification

Huy Nguyen, Kien Nguyen, Akila Pemasiri, Feng Liu, Sridha Sridharan, Clinton Fookes; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1241-1251

We introduce AG-VPReID, a new large-scale dataset for aerial-ground video-based person re-identification (ReID) that comprises 6,632 subjects, 32,321 tracklets and over 9.6 million frames captured by drones (altitudes ranging from 15-120m), CCTV, and wearable cameras. This dataset offers a real-world benchmark for evaluating the robustness to significant viewpoint changes, scale variations, and resolution differences in cross-platform aerial-ground settings. In addition, to address these challenges, we propose AG-VPReID-Net, an end-to-end framework composed of three complementary streams: (1) an Adapted Temporal-Spatial Stream addressing motion pattern inconsistencies and facilitating temporal feature learning, (2) a Normalized Appearance Stream leveraging physics-informed techniques to tackle resolution and appearance changes, and (3) a Multi-Scale Attention Stream handling scale variations across drone altitudes. We integrate visual-semantic cues from all streams to form a robust, viewpoint-invariant whole-body representation. Extensive experiments demonstrate that AG-VPReID-Net outperforms state-of-the-art approaches on both our new dataset and existing video-based ReID benchmarks, showcasing its effectiveness and generalizability. Nevertheless, the performance gap observed on AG-VPReID across all methods underscores the dataset's challenging nature. The dataset, code and trained models are available at https://github.com/agvpreid25/AG-VPReID-Net.

****

EBS-EKF: Accurate and High Frequency Event-based Star Tracking

Albert W. Reed, Connor Hashemi, Dennis Melamed, Nitesh Menon, Keigo Hirakawa, Scott McCloskey; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6510-6519

Event-based sensors (EBS) are a promising new technology for star tracking due to their low latency and power efficiency, but prior work has thus far been evaluated exclusively in simulation with simplified signal models. We propose a novel algorithm for event-based star tracking, grounded in an analysis of the EBS circuit and an extended Kalman filter (EKF). We quantitatively evaluate our method using real night sky data, comparing its results with those from a space-ready active-pixel sensor (APS) star tracker. We demonstrate that our method is an order-of-magnitude more accurate than existing methods due to improved signal modeling and state estimation, while providing more frequent updates and greater motion tolerance than conventional APS trackers. We provide all code and the first dataset of events synchronized with APS solutions.

****

PersonaBooth: Personalized Text-to-Motion Generation

Boeun Kim, Hea In Jeong, JungHoon Sung, Yihua Cheng, Jeongmin Lee, Ju Yong Chang, Sang-Il Choi, Younggeun Choi, Saim Shin, Jungho Kim, Hyung Jin Chang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22756-22765

This paper introduces Motion Personalization, a new task that generates personalized motions aligned with text descriptions using several basic motions containing Persona. To support this novel task, we introduce a new large-scale motion dataset called PerMo (PersonaMotion), which captures the unique personas of multiple actors. We also propose a multi-modal finetuning method of a pretrained motion diffusion model called PersonaBooth. PersonaBooth addresses two main challenges: i) A significant distribution gap between the persona-focused PerMo dataset and the pretraining datasets, which lack persona-specific data, and ii) the difficulty of capturing a consistent persona from the motions vary in content (action type). To tackle the dataset distribution gap, we introduce a persona token to accept new persona features and perform multi-modal adaptation for both text and visuals during finetuning. To capture a consistent persona, we incorporate a contrastive learning technique to enhance intra-cohesion among samples with the same persona. Furthermore, we introduce a context-aware fusion mechanism to maximize the integration of persona cues from multiple input motions. PersonaBooth outperforms state-of-the-art motion style transfer methods, establishing a new benchmark for motion personalization.

****

Benchmarking Object Detectors under Real-World Distribution Shifts in Satellite

Imagery
Sara A. Al-Emadi, Yin Yang, Ferda Ofli; Proceedings of the Computer Vision and P attern Recognition Conference (CVPR), 2025, pp. 8299-8309
Object detectors have achieved remarkable performance in many applications; howe ver, these deep learning models are typically designed under the i.i.d. assumpti on, meaning they are trained and evaluated on data sampled from the same (source) distribution. In real-world deployment, however, target distributions often di ffer from source data, leading to substantial performance degradation. Domain Ge neralisation (DG) seeks to bridge this gap by enabling models to generalise to O ut-Of-Distribution (OOD) data without access to target distributions during trai ning, enhancing robustness to unseen conditions. In this work, we examine the ge neralisability and robustness of state-of-the-art object detectors under real-wo rld distribution shifts, focusing particularly on spatial domain shifts. Despite the need, a standardised benchmark dataset specifically designed for assessing object detection under realistic DG scenarios is currently lacking. To address t his, we introduce Real-World Distribution Shifts (RWDS), a suite of three novel DG benchmarking datasets that focus on humanitarian and climate change applicati ons. These datasets enable the investigation of domain shifts across (i) climate zones and (ii) various disasters and geographic regions. To our knowledge, thes e are the first DG benchmarking datasets tailored for object detection in real-w orld, high-impact contexts. We aim for these datasets to serve as valuable resou rces for evaluating the robustness and generalisation of future object detection models. Our datasets and code are available at https://github.com/RWGAI/RWDS.
****

SAIST: Segment Any Infrared Small Target Model Guided by Contrastive Language-Im age Pretraining
Mingjin Zhang, Xiaolong Li, Fei Gao, Jie Guo, Xinbo Gao, Jing Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 954 9-9558
Infrared Small Target Detection (IRSTD) aims to identify low signal-to-noise rat io small targets in infrared images with complex backgrounds, which is crucial f or various applications. However, existing IRSTD methods typically rely solely o n image modalities for processing, which fail to fully capture contextual inform ation, leading to limited detection accuracy and adaptability in complex environ ments. Inspired by vision-language models, this paper proposes a novel framework , SAIST, which integrates textual information with image modalities to enhance I RSTD performance. The framework consists of two main components: Scene Recogniti on Contrastive Language-Image Pretraining (SR-CLIP) and CLIP-guided Segment Anyt hing Model (CG-SAM). SR-CLIP generates a set of visual descriptions through obj ect-object similarity and object-scene relevance, embedding them into learnable prompts to refine the textual description set. This reduces the domain gap betwe en vision and language, generating precise textual and visual prompts. CG-SAM ut ilizes the prompts generated by SR-CLIP to accurately guide the Mask Decoder in learning prior knowledge of background features, while incorporating infrared im aging equations to improve small target recognition in complex backgrounds and s ignificantly reduce the false alarm rate. Additionally, this paper introduces th e first multimodal IRSTD dataset, MIRSTD, which contains abundant image-text pai rs. Experimental results demonstrate that the proposed SAIST method outperforms existing state-of-the-art approaches.
****

Star with Bilinear Mapping
Zelin Peng, Yu Huang, Zhengqin Xu, Feilong Tang, Ming Hu, Xiaokang Yang, Wei She n; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25292-25302
Contextual modeling is crucial for robust visual representation learning, especi ally in computer vision. Although Transformers have become a leading architectur e for vision tasks due to their attention mechanism, the quadratic complexity of full attention operations presents substantial computational challenges. To add ress this, we introduce Star with Bilinear Mapping (SBM), a Transformer-like arc hitecture that achieves global contextual modeling with linear complexity. SBM e

mploys a bilinear mapping module (BM) with low-rank decomposition strategy and s tar operations (element-wise multiplication) to efficiently capture global conte xtual information. Our model demonstrates competitive performance on image class ification and semantic segmentation tasks, delivering significant computational efficiency gains compared to traditional attention-based models.
****

Closed-Loop Supervised Fine-Tuning of Tokenized Traffic Models
Zhejun Zhang, Peter Karkus, Maximilian Igl, Wenhao Ding, Yuxiao Chen, Boris Ivan ovic, Marco Pavone; Proceedings of the Computer Vision and Pattern Recognition C onference (CVPR), 2025, pp. 5422-5432
Traffic simulation aims to learn a policy for traffic agents that, when unrolled in closed-loop, faithfully recovers the joint distribution of trajectories obse rved in the real world. Inspired by large language models, tokenized multi-agent policies have recently become the state-of-the-art in traffic simulation. Howev er, they are typically trained through open-loop behavior cloning, and thus suff er from covariate shift when executed in closed-loop during simulation. In this work, we present Closest Among Top-K (CAT-K) rollouts, a simple yet effective cl osed-loop fine-tuning strategy to mitigate covariate shift. CAT-K fine-tuning on ly requires existing trajectory data, without reinforcement learning or generati ve adversarial imitation. Concretely, CAT-K fine-tuning enables a small 7M-param eter tokenized traffic simulation policy to outperform a 102M-parameter model fr om the same model family, achieving the top spot on the Waymo Sim Agent Challeng e leaderboard at the time of submission. The code is available at https://github .com/NVlabs/catk.
****

DIFIX3D+: Improving 3D Reconstructions with Single-Step Diffusion Models
Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng S hou, Sanja Fidler, Zan Gojcic, Huan Ling; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26024-26035
Neural Radiance Fields and 3D Gaussian Splatting have revolutionized 3D reconstr uction and novel-view synthesis task. However, achieving photorealistic renderin g from extreme novel viewpoints remains challenging, as artifacts persist across representations. In this work, we introduce Difix3D+, a novel pipeline designed to enhance 3D reconstruction and novel-view synthesis through single-step diffu sion models. At the core of our approach is Difix, a single-step image diffusion model trained to enhance and remove artifacts in rendered novel views caused by underconstrained regions of the 3D representation.Difix serves two critical rol es in our pipeline. First, it is used during the reconstruction phase to clean u p pseudo-training views that are rendered from the reconstruction and then disti lled back into 3D. This greatly enhances underconstrained regions and improves t he overall 3D representation quality. More importantly, Difix also acts as a neu ral enhancer during inference, effectively removing residual artifacts arising f rom imperfect 3D supervision and the limited capacity of current reconstruction models. Difix3D+ is a general solution, a single model compatible with both NeRF and 3DGS representations, and it achieves an average 2x improvement in FID scor e over baselines while maintaining 3D consistency.
****

Time of the Flight of the Gaussians: Optimizing Depth Indirectly in Dynamic Radi ance Fields
Runfeng Li, Mikhail Okunev, Zixuan Guo, Anh Ha Duong, Christian Richardt, Matthe w O'Toole, James Tompkin; Proceedings of the Computer Vision and Pattern Recogni tion Conference (CVPR), 2025, pp. 21021-21030
We present a method to reconstruct dynamic scenes from monocular continuous-wave time-of-flight (C-ToF) cameras using raw sensor samples that achieves similar o r better accuracy than neural volumetric approaches and is 100xfaster. Quickly a chieving high-fidelity dynamic 3D reconstruction from a single viewpoint is a si gnificant challenge in computer vision. In C-ToF radiance field reconstruction, the property of interest---depth---is not directly measured, causing an addition al challenge. This problem has a large and underappreciated impact upon the opti mization when using a fast primitive-based scene representation like 3D Gaussian

splatting, which is commonly used with multi-view data to produce satisfactory results and is brittle in its optimization otherwise. We incorporate two heuristics into the optimization to improve the accuracy of scene geometry represented by Gaussians. Experimental results show that our approach produces accurate reconstructions under constrained C-ToF sensing conditions, including for fast motions like swinging baseball bats.

****

Align3R: Aligned Monocular Depth Estimation for Dynamic Videos

Jiahao Lu, Tianyu Huang, Peng Li, Zhiyang Dou, Cheng Lin, Zhiming Cui, Zhen Dong, Sai-Kit Yeung, Wenping Wang, Yuan Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22820-22830

Recent developments in monocular depth estimation methods enable high-quality depth estimation of single-view images but fail to estimate consistent video depth across different frames. Recent works address this problem by applying a video diffusion model to generate video depth conditioned on the input video, which is training-expensive and can only produce scale-invariant depth values without camera poses. In this paper, we propose a novel video-depth estimation method called Align3R to estimate temporal consistent depth maps for a dynamic video. Our key idea is to utilize the recent DUSt3R model to align estimated monocular depth maps of different timesteps. First, we fine-tune the DUSt3R model with additional estimated monocular depth as inputs for the dynamic scenes. Then, we apply optimization to reconstruct both depth maps and camera poses. Extensive experiments demonstrate that Align3R estimates consistent video depth and camera poses for a monocular video with superior performance than baseline methods.

****

Compositional Caching for Training-free Open-vocabulary Attribute Detection

Marco Garosi, Alessandro Conti, Gaowen Liu, Elisa Ricci, Massimiliano Mancini; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15098-15107

Attribute detection is crucial for many computer vision tasks, as it enables systems to describe properties such as color, texture, and material. Current approaches often rely on labor-intensive annotation processes which are inherently limited: objects can be described at an arbitrary level of detail (e.g., color vs. color shades), leading to ambiguities when the annotators are not instructed carefully. Furthermore, they operate within a predefined set of attributes, reducing scalability and adaptability to unforeseen downstream applications. We present Compositional Caching (ComCa), a training-free method for open-vocabulary attribute detection that overcomes these constraints. ComCa requires only the list of target attributes and objects as input, using them to populate an auxiliary cache of images by leveraging web-scale databases and Large Language Models to determine attribute-object compatibility. To account for the compositional nature of attributes, cache images receive soft attribute labels. Those are aggregated at inference time based on the similarity between the input and cache images, refining the predictions of underlying Vision-Language Models (VLMs). Importantly, our approach is model-agnostic, compatible with various VLMs. Experiments on public datasets demonstrate that ComCa significantly outperforms zero-shot and cache-based baselines, competing with recent training-based methods, proving that a carefully designed training-free approach can successfully address open-vocabulary attribute detection.

****

Seek Common Ground While Reserving Differences: Semi-Supervised Image-Text Sentiment Recognition

Wuyou Xia, Guoli Jia, Sicheng Zhao, Jufeng Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29601-29611

Multimodal sentiment analysis has attracted extensive research attention as increasing users share images and texts to express their emotions and opinions on social media. Collecting large amounts of labeled sentiment data is an expensive and challenging task due to the high cost of labeling and unavoidable label ambiguity. Semi-supervised learning (SSL) is explored to utilize the extensive unlabeled data to alleviate the demand for annotation. However, different from typical

multimodal tasks, the inconsistent sentiment between image and text leads to th
e sub-optimal performance of SSL algorithms. To address this issue, we propose S
CDR, the first semi-supervised image-text sentiment recognition framework. To be
tter utilize the discriminative features of each modality, we decouple features
into common and private parts and then use the private features to train unimoda
l classifiers for enhanced modality-specific sentiment representation. Consideri
ng the complex relation between modalities, we devise a modal selection-based at
tention module that adaptively assesses the dominant sentiment modality at the s
ample level to guide the fusion of multimodal representations. Furthermore, to p
revent the model predictions from overly relying on common features under the gu
idance of multimodal labels, we design a pseudo-label filtering strategy based o
n the matching degree of prediction and dominant modality. Extensive experiments
and comparisons on five publicly available datasets demonstrate that SCDR outpe
rforms state-of-the-art methods. The code is provided in the supplementary mater
ial and will be released to the public.
****

CoLLM: A Large Language Model for Composed Image Retrieval
Chuong Huynh, Jinyu Yang, Ashish Tawari, Mubarak Shah, Son Tran, Raffay Hamid, T
rishul Chilimbi, Abhinav Shrivastava; Proceedings of the Computer Vision and Pat
tern Recognition Conference (CVPR), 2025, pp. 3994-4004
Composed Image Retrieval (CIR) is a complex task that aims to retrieve images ba
sed on a multimodal query. Typical training data consists of triplets containing
a reference image, a textual description of desired modifications, and the targ
et image, which are expensive and time-consuming to acquire. The scarcity of CIR
datasets has led to zero-shot approaches utilizing synthetic triplets or levera
ging vision-language models (VLMs) with ubiquitous web-crawled image-caption pai
rs. However, these methods have significant limitations: synthetic triplets suff
er from limited scale, lack of diversity, and unnatural modification text, while
image-caption pairs hinder joint embedding learning of the multimodal query due
to the absence of triplet data. Moreover, existing approaches struggle with com
plex and nuanced modification texts that demand sophisticated fusion and underst
anding of vision and language modalities. We present CoLLM, a one-stop framework
that effectively addresses these limitations. Our approach generates triplets o
n-the-fly from image-caption pairs, enabling supervised training without manual
annotation. We leverage Large Language Models (LLMs) to generate joint embedding
s of reference images and modification texts, facilitating deeper multimodal fus
ion. Additionally, we introduce Multi-Text CIR (MTCIR), a large-scale dataset co
mprising 3.4M samples, and refine existing CIR benchmarks (CIRR and Fashion-IQ)
to enhance evaluation reliability. Experimental results demonstrate that CoLLM a
chieves state-of-the-art performance across multiple CIR benchmarks and settings
. MTCIR yields competitive results, with up to 15% performance improvement. Our
refined benchmarks provide more reliable evaluation metrics for CIR models, cont
ributing to the advancement of this important field. Project page is at collm-cv
pr25.github.io
****

Anomize: Better Open Vocabulary Video Anomaly Detection
Fei Li, Wenxuan Liu, Jingjing Chen, Ruixu Zhang, Yuran Wang, Xian Zhong, Zheng W
ang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR
), 2025, pp. 29203-29212
Open Vocabulary Video Anomaly Detection (OVVAD) seeks to detect and classify bot
h base and novel anomalies. However, existing methods face two specific challeng
es related to novel anomalies. The first challenge is detection ambiguity, where
the model struggles to assign accurate anomaly scores to unfamiliar anomalies.
The second challenge is categorization confusion, where novel anomalies are ofte
n misclassified as visually similar base instances. To address these challenges,
we explore supplementary information from multiple sources to mitigate detectio
n ambiguity by leveraging multiple levels of visual data alongside matching text
ual information. Furthermore, we propose incorporating label relations to guide
the encoding of new labels, thereby improving alignment between novel videos and
their corresponding labels, which helps reduce categorization confusion. The re

sulting Anomize framework effectively tackles these issues, achieving superior performance on UCF-Crime and XD-Violence datasets, demonstrating its effectiveness in OVVAD.
****

Efficient Diffusion as Low Light Enhancer
Guanzhou Lan, Qianli Ma, Yuqi Yang, Zhigang Wang, Dong Wang, Xuelong Li, Bin Zhao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21277-21286
The computational burden of the iterative sampling process remains a major challenge in diffusion-based Low-Light Image Enhancement (LLIE). Current acceleration methods, whether training-based or training-free, often lead to significant performance degradation, highlighting the trade-off between performance and efficiency.In this paper, we identify two primary factors contributing to performance degradation: fitting errors and the inference gap. Our key insight is that fitting errors can be mitigated by linearly extrapolating the incorrect score functions, while the inference gap can be reduced by shifting the Gaussian flow to a reflectance-aware residual space.Based on the above insights, we design Reflectance-Aware Trajectory Refinement (RATR) module, a simple yet effective module to refine the teacher trajectory using the reflectance component of images. Following this, we introduce Reflectance-aware Diffusion with Distilled Trajectory ReDDiT, an efficient and flexible distillation framework tailored for LLIE. Our framework achieves comparable performance to previous diffusion-based methods with redundant steps in just 2 steps while establishing new state-of-the-art (SOTA) results with 8 or 4 steps. Comprehensive experimental evaluations on 10 benchmark datasets validate the effectiveness of our method, consistently outperforming existing SOTA methods.
****

GraphMimic: Graph-to-Graphs Generative Modeling from Videos for Policy Learning
Guangyan Chen, Te Cui, Meiling Wang, Chengcai Yang, Mengxiao Hu, Haoyang Lu, Yao Mu, Zicai Peng, Tianxing Zhou, Xinran Jiang, Yi Yang, Yufeng Yue; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1756-1768
Learning from demonstration is a powerful method for robotic skill acquisition. However, the significant expense of collecting such action-labeled robot data presents a major bottleneck. Video data, a rich data source encompassing diverse behavioral and physical knowledge, emerges as a promising alternative. In this paper, we present GraphMimic, a novel paradigm that leverages video data via graph-to-graphs generative modeling, which pre-trains models to generate future graphs conditioned on the graph within a video frame. Specifically, GraphMimic abstracts video frames into object and visual action vertices, and constructs graphs for state representations. The graph generative modeling network then effectively models internal structures and spatial relationships within the constructed graphs, aiming to generate future graphs. The generated graphs serve as conditions for the control policy, mapping to robot actions. Our concise approach captures important spatial relations and enhances future graph generation accuracy, enabling the acquisition of robust policies from limited action-labeled data. Furthermore, the transferable graph representations facilitate the effective learning of manipulation skills from cross-embodiment videos. Our experiments exhibit that GraphMimic achieves superior performance using merely 20% action-labeled data. Moreover, our method outperforms the state-of-the-art method by over 17% and 23% in simulation and real-world experiments, and delivers improvements of over 33% in cross-embodiment transfer experiments.
****

VI^3NR: Variance Informed Initialization for Implicit Neural Representations
Chamin Hewa Koneputugodage, Yizhak Ben-Shabat, Sameera Ramasinghe, Stephen Gould; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13477-13486
Implicit Neural Representations (INRs) are a versatile and powerful tool for encoding various forms of data, including images, videos, sound, and 3D shapes. A critical factor in the success of INRs is the initialization of the network, whic

h can significantly impact the convergence and accuracy of the learned model. Unfortunately, commonly used neural network initializations are not widely applicable for many activation functions, especially those used by INRs. In this paper, we improve upon previous initialization methods by deriving an initialization that has stable variance across layers, and applies to any activation function. We show that this generalizes many previous initialization methods, and has even better stability for well studied activations. We also show that our initialization leads to improved results with INR activation functions in multiple signal modalities. Our approach is particularly effective for Gaussian INRs, where we demonstrate that the theory of our initialization matches with task performance in multiple experiments, allowing us to achieve improvements in image, audio, and 3D surface reconstruction.

****

MMVU: Measuring Expert-Level Multi-Discipline Video Understanding

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, Chengye Wang, Ziyao Shangguan, Zhenwen Liang, Yixin Liu, Chen Zhao, Arman Cohan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8475-8489

We introduce MMVU, a comprehensive expert-level, multi-discipline benchmark for evaluating foundation models in video understanding. MMVU includes 3,000 expert-annotated questions spanning 27 subjects across four core disciplines: Science, Healthcare, Humanities & Social Sciences, and Engineering. Compared to prior benchmarks, MMVU features three key advancements. First, it challenges models to apply domain-specific knowledge and perform expert-level reasoning to analyze specialized-domain videos, moving beyond the basic visual perception typically assessed in current video benchmarks. Second, each example is annotated by human experts from scratch. We implement strict data quality controls to ensure the high quality of the dataset. Finally, each example is enriched with expert-annotated reasoning rationals and relevant domain knowledge, facilitating in-depth analysis. We conduct an extensive evaluation of 36 frontier multimodal foundation models on MMVU. The latest System-2-capable models, o1 and Gemini 2.0 Flash Thinking, achieve the highest performance among the tested models. However, they still fall short of matching human expertise. Through in-depth error analyses and case studies, we offer actionable insights for future advancements in expert-level, knowledge-intensive video understanding for specialized domains.

****

M-LLM Based Video Frame Selection for Efficient Video Understanding

Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, Trishul Chilimbi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13702-13712

Recent advances in Multi-Modal Large Language Models (M-LLMs) show promising results in video reasoning. Popular Multi-Modal Large Language Model (M-LLM) frameworks usually apply naive uniform sampling to reduce the number of video frames that are fed into an M-LLM, particularly for long context videos. However, it could lose crucial context in certain periods of a video, so that the downstream M-LLM may not have sufficient visual information to answer a question. To attack this pain point, we propose a light-weight M-LLM -based frame selection method that adaptively select frames that are more relevant to users' queries. In order to train the proposed frame selector, we introduce two supervision signals (i) Spatial signal, where single frame importance score by prompting an M-LLM; (ii) Temporal signal, in which multiple frames selection by prompting Large Language Model (LLM) using the captions of all frame candidates. The selected frames are then digested by a frozen downstream video M-LLM for visual reasoning and question answering. Empirical results show that the proposed M-LLM video frame selector improves the performances various downstream video Large Language Model (video-LLM) across medium (ActivityNet, NExT-QA) and long (EgoSchema, LongVideoBench) context video question answering benchmarks.

****

Search and Detect: Training-Free Long Tail Object Detection via Web-Image Retrieval

Mankeerat Sidhu, Hetarth Chopra, Ansel Blume, Jeonghwan Kim, Revanth Gangi Reddy, Heng Ji; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15129-15138
In this paper, we introduce SearchDet, a training-free long-tail object detection framework that significantly enhances open-vocabulary object detection performance. SearchDet retrieves a set of positive and negative images of an object to ground, embeds these images, and computes an input image--weighted query which is used to detect the desired concept in the image. Our proposed method is simple and training-free, yet achieves over 16.81% mAP improvement on ODinW and 59.85% mAP improvement on LVIS compared to state-of-the-art models such as GroundingDINO. We further show that our approach of basing object detection on a set of Web-retrieved exemplars is stable with respect to variations in the exemplars, suggesting a path towards eliminating costly data annotation and training procedures.
****
EgoLM: Multi-Modal Language Model of Egocentric Motions
Fangzhou Hong, Vladimir Guzov, Hyo Jin Kim, Yuting Ye, Richard Newcombe, Ziwei Liu, Lingni Ma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5344-5354
As wearable devices become more prevalent, understanding the user's motion is crucial for improving contextual AI systems. We introduce EgoLM, a versatile framework designed for egocentric motion understanding using multi-modal data. EgoLM integrates the rich contextual information from egocentric videos and motion sensors afforded by wearable devices. It also combines dense supervision signals from motion and language, leveraging the vast knowledge encoded in pre-trained large language models (LLMs). EgoLM models the joint distribution of egocentric motions and natural language using LLMs, conditioned on observations from egocentric videos and motion sensors. It unifies a range of motion understanding tasks, including motion narration from video or motion data, as well as motion generation from text or sparse sensor data. Unique to wearable devices, it also enables a novel task to generate text descriptions from sparse sensors. Through extensive experiments, we validate the effectiveness of EgoLM in addressing the challenges of under-constrained egocentric motion learning, and demonstrate its capability as a generalist model through a variety of applications.
****
Unleashing the Potential of Multi-modal Foundation Models and Video Diffusion for 4D Dynamic Physical Scene Simulation
Zhuoman Liu, Weicai Ye, Yan Luximon, Pengfei Wan, Di Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11016-11025
Realistic simulation of dynamic scenes requires accurately capturing diverse material properties and modeling complex object interactions grounded in physical principles. However, existing methods are constrained to basic material types with limited predictable parameters, making them insufficient to represent the complexity of real-world materials. We introduce PhysFlow, a novel approach that leverages multi-modal foundation models and video diffusion to achieve enhanced 4D dynamic scene simulation. Our method utilizes multi-modal models to identify material types and initialize material parameters through image queries, while simultaneously inferring 3D Gaussian splats for detailed scene representation. We further refine these material parameters using video diffusion with a differentiable Material Point Method (MPM) and optical flow guidance rather than render loss or Score Distillation Sampling (SDS) loss. This integrated framework enables accurate prediction and realistic simulation of dynamic interactions in real-world scenarios, advancing both accuracy and flexibility in physics-based simulations.
****
Diffusion Model is Effectively Its Own Teacher
Xinyin Ma, Runpeng Yu, Songhua Liu, Gongfan Fang, Xinchao Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12901-12911
In this paper, we introduce a novel self-distillation paradigm for improving the

performance of diffusion models. Previous studies have shown that introducing a teacher to distill the diffusion model can enhance its sampling efficiency. We raise an intriguing question: can the diffusion model itself serve as its teacher to further improve the performance of itself? To this end, we propose a new paradigm called Self Step-Distillation (SSD). The core idea of SSD is to integrate the predictions or the intermediate activations of the diffusion model at each timestep with its preceding timestep through a fusion mechanism. We propose two forms, explicit SSD and implicit SSD (iSSD), to perform N-step to N-step distillation from the diffusion model itself to achieve improved image quality. We further elucidate the relationship between SSD and high-order solver, highlighting their underlying relationship. The effectiveness of SSD is validated through extensive experiments on diffusion transformers of various sizes and across different sampling steps. Our results show that this novel self-distillation paradigm can significantly enhance performance. Additionally, our method is compatible with the distillation method designed for few-step inference. Notably, with iSSD trained less than one epoch, we obtain a 32-step DiT-XL/2 achieving an FID of 1.99, outperforming the original 250-step DiT-XL/2 with an FID of 2.26. We further validate the effectiveness of our method on text-to-image diffusion models, such as Stable Diffusion, and also observe notable improvement in image quality.

****

Long Video Diffusion Generation with Segmented Cross-Attention and Content-Rich Video Data Curation

Xin Yan, Yuxuan Cai, Qiuyue Wang, Yuan Zhou, Wenhao Huang, Huan Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3184-3194

We introduce Presto, a novel video diffusion model designed to generate 15-second videos with long-range coherence and rich content. Extending video generation to maintain scenario diversity over long durations presents significant challenges. To address this, we propose a Segmented Cross-Attention (SCA) strategy, which splits hidden states into segments along the temporal dimension, allowing each segment to cross-attend to a corresponding sub-caption. SCA requires no additional parameters, enabling seamless incorporation into current DiT-based architectures. To facilitate high-quality long video generation, we build the LongTake-HD dataset, consisting of 261k content-rich videos with scenario coherence, annotated with an overall video caption and five progressive sub-captions. Experiments show that our Presto achieves 78.5% on the VBench Semantic Score and 100% on the Dynamic Degree, outperforming existing state-of-the-art video generation methods. This demonstrates that our proposed Presto significantly enhances the content richness, maintains long-range coherence, and captures intricate textual details. All code and model weights will be made publicly available.

****

Learning Heterogeneous Tissues with Mixture of Experts for Gigapixel Whole Slide Images

Junxian Wu, Minheng Chen, Xinyi Ke, Tianwang Xun, Xiaoming Jiang, Hongyu Zhou, Lizhi Shao, Youyong Kong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5144-5153

Analyzing gigapixel Whole Slide Images (WSIs) is challenging due to the complex pathological tissue environment and the absence of target-driven domain knowledge. Previous methods incorporated pathological priors to mitigate this issue but relied on additional inference steps and specialized workflows, restricting scalability and the model's capacity to identify novel outcome-related factors. To address these challenges, we propose a plug-and-play Pathology-Aware Mixture-of-Experts (PAMoE) module, which based on mixture of experts to learn pathology-related knowledge and extract useful information. We train the experts to become 'specialists' in specific intratumoral tissues by learning to route each tissue to its mapped expert. In addition, to reduce the impact of irrelevant content on the model, we introduce a new routing rule that discards patches in which none of the experts express interest, which helps the model better capture the relationships between relevant patches. Through a comprehensive evaluation of PAMoE on survival task, we demonstrate that 1) Our module enhances the performance of basel

ine models in most cases, and 2) The sparse expert processing across different t issues enhances the learning of patch representations by addressing tissue heter ogeneity.
****

HyperNVD: Accelerating Neural Video Decomposition via Hypernetworks
Maria Pilligua, Danna Xue, Javier Vazquez-Corral; Proceedings of the Computer Vi sion and Pattern Recognition Conference (CVPR), 2025, pp. 22933-22942
Decomposing a video into a layer-based representation is crucial for easy video editing for the creative industries, as it enables independent editing of specif ic layers. Existing video-layer decomposition models rely on implicit neural rep resentations (INRs) trained independently for each video, making the process tim e-consuming when applied to new videos. Noticing this limitation, we propose a m eta-learning strategy to learn a generic video decomposition model to speed up t he training on new videos. Our model is based on a hypernetwork architecture whi ch, given a video-encoder embedding, generates the parameters for a compact INR- based neural video decomposition model. Our strategy mitigates the problem of si ngle-video overfitting and, importantly, shortens the convergence of video decom position on new, unseen videos. Our code is available at: https://hypernvd.githu b.io/
****

UnCommon Objects in 3D
Xingchen Liu, Piyush Tayal, Jianyuan Wang, Jesus Zarzar, Tom Monnier, Konstantin os Tertikas, Jiali Duan, Antoine Toisoul, Jason Y. Zhang, Natalia Neverova, Andr ea Vedaldi, Roman Shapovalov, David Novotny; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14102-14113
We introduce Uncommon Objects in 3D (uCO3D), a new object-centric dataset for 3D deep learning and 3D generative AI. uCO3D is the largest publicly-available col lection of high-resolution videos of objects with 3D annotations that ensures fu ll-360 degree coverage. uCO3D is significantly more diverse than MVImgNet and CO 3Dv2, covering more than 1,000 object categories. It is also of higher quality, due to extensive quality checks of both the collected videos and the 3D annotati ons. Similar to analogous datasets, uCO3D contains annotations for 3D camera pos es, depth maps and sparse point clouds. In addition, each object is equipped wit h a caption and a 3D Gaussian Splat reconstruction. We train several large 3D mo dels on MVImgNet, CO3Dv2, and uCO3D and obtain superior results using the latter , showing that uCO3D is better for learning applications.
****

Disentangled Pose and Appearance Guidance for Multi-Pose Generation
Tengfei Xiao, Yue Wu, Yuelong Li, Can Qin, Maoguo Gong, Qiguang Miao, Wenping Ma ; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5646-5655
Human pose generation is a complex task due to the non-rigid and highly variable nature of human body structures and appearances. However, existing methods ofte n overlook the fundamental differences between spatial transformations of poses and texture generation for appearance, which makes them prone to overfitting. To address this issue, we propose a multi-pose generation framework driven by dise ntangled pose and appearance guidance. Our approach includes a Global-aware Pose Generation module that iteratively generates pose embeddings, enabling effectiv e control over non-rigid body deformations. Additionally, we introduce the Globa l-aware Transformer Decoder, which leverages similarity queries and attention me chanisms to achieve spatial transformations and enhance pose consistency through a Global-aware block. In the appearance generation phase, we condition a diffus ion model on pose embeddings produced in the initial stage and introduce an Appe arance Adapter that extracts high-level contextual semantic information from mul ti-scale features, enabling further refinement of pose appearance textures and p roviding appearance guidance. Extensive experiments on the UBC Fashion and TikTo k datasets demonstrate that our framework achieves state-of-the-art results in b oth quality and fidelity, establishing it as a powerful approach for complex pos e generation tasks.
****

Mind the Gap: Confidence Discrepancy Can Guide Federated Semi-Supervised Learning Across Pseudo-Mismatch
Yijie Liu, Xinyi Shang, Yiqun Zhang, Yang Lu, Chen Gong, Jing-Hao Xue, Hanzi Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10173-10182
Federated Semi-Supervised Learning (FSSL) aims to leverage unlabeled data across clients with limited labeled data to train a global model with strong generalization ability. Most FSSL methods rely on consistency regularization with pseudo-labels, converting predictions from local or global models into hard pseudo-labels as supervisory signals. However, we discover that the quality of pseudo-label is largely deteriorated by data heterogeneity, an intrinsic facet of federated learning. In this paper, we study the problem of FSSL in-depth and show that (1) heterogeneity exacerbates pseudo-label mismatches, further degrading model performance and convergence, and (2) local and global models' predictive tendencies diverge as heterogeneity increases. Motivated by these findings, we propose a simple and effective method called Semi-supervised Aggregation for Globally-Enhanced Ensemble (SAGE), that can flexibly correct pseudo-labels based on confidence discrepancies. This strategy effectively mitigates performance degradation caused by incorrect pseudo-labels and enhances consensus between local and global models. Experimental results demonstrate that SAGE outperforms existing FSSL methods in both performance and convergence. Our code is available at \href https://github.com/Jay-Codeman/SAGE https://github.com/Jay-Codeman/SAGE .
****

Instant Adversarial Purification with Adversarial Consistency Distillation
Chun Tong Lei, Hon Ming Yam, Zhongliang Guo, Yifei Qian, Chun Pong Lau; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24331-24340
Neural networks have revolutionized numerous fields with their exceptional performance, yet they remain susceptible to adversarial attacks through subtle perturbations. While diffusion-based purification methods like DiffPure offer promising defense mechanisms, their computational overhead presents a significant practical limitation. In this paper, we introduce One Step Control Purification (OSCP), a novel defense framework that achieves robust adversarial purification in a single Neural Function Evaluation (NFE) within diffusion models. We propose Gaussian Adversarial Noise Distillation (GAND) as the distillation objective and Controlled Adversarial Purification (CAP) as the inference pipeline, which makes OSCP demonstrate remarkable efficiency while maintaining defense efficacy. Our proposed GAND addresses a fundamental tension between consistency distillation and adversarial perturbation, bridging the gap between natural and adversarial manifolds in the latent space, while remaining computationally efficient through Parameter-Efficient Fine-Tuning (PEFT) methods such as LoRA, eliminating the high computational budget request from full parameter fine-tuning. The CAP guides the purification process through the unlearnable edge detection operator calculated by the input image as an extra prompt, effectively preventing the purified images from deviating from their original appearance when using large purification steps. Our experimental results on ImageNet showcase OSCP's superior performance, achieving a 74.19% defense success rate with merely 0.1s per purification --- a 100-fold speedup compared to conventional approaches.
****

Learning Textual Prompts for Open-World Semi-Supervised Learning
Yuxin Fan, Junbiao Cui, Jiye Liang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14756-14765
Traditional semi-supervised learning achieves significant success in closed-world scenarios. To better align with the openness of the real world, researchers propose open-world semi-supervised learning (OWSSL), which enables models to effectively recognize known and unknown classes even without labels for unknown classes. Recently, researchers have attempted to enhance the model performance in recognizing visually similar classes by integrating textual information. However, these attempts do not effectively align images with text, resulting in limited improvements in model performance. In response to this challenge, we propose a nov

el OWSSL method. By adopting a global-and-local textual prompt learning strategy to enhance image-text alignment effectiveness, and implementing a forward-and-backward strategy to reduce noise in image-text matching for unlabeled samples, we ultimately enhance the model's ability to extract and recognize discriminative features across different classes. Experimental results on multiple fine-grained datasets demonstrate that our method achieves significant performance improvements compared to state-of-the-art methods.
****

Electromyography-Informed Facial Expression Reconstruction for Physiological-Based Synthesis and Analysis
Tim Büchner, Christoph Anders, Orlando Guntinas-Lichius, Joachim Denzler; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 215-227
The relationship between muscle activity and resulting facial expressions is crucial for various fields, including psychology, medicine, and entertainment. The synchronous recording of facial mimicry and muscular activity via surface electromyography (sEMG) provides a unique window into these complex dynamics. Unfortunately, existing methods for facial analysis cannot handle electrode occlusion, rendering them ineffective. Even with occlusion-free reference images of the same person, variations in expression intensity and execution are unmatchable. Our electromyography-informed facial expression reconstruction (EIFER) approach is a novel method to restore faces under sEMG occlusion faithfully in an adversarial manner. We decouple facial geometry and visual appearance (e.g., skin texture, lighting, electrodes) by combining a 3D Morphable Model (3DMM) with neural unpaired image-to-image translation via reference recordings. Then, EIFER learns a bidirectional mapping between 3DMM expression parameters and muscle activity, establishing correspondence between the two domains. We validate the effectiveness of our approach through experiments on a dataset of synchronized sEMG recordings and facial mimicry, demonstrating faithful geometry and appearance reconstruction. Further, we synthesize expressions based on muscle activity and how observed expressions can predict dynamic muscle activity. Consequently, EIFER introduces a new paradigm for facial electromyography, which could be extended to other forms of multi-modal face recordings.
****

LongDiff: Training-Free Long Video Generation in One Go
Zhuoling Li, Hossein Rahmani, Qiuhong Ke, Jun Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17789-17798
Video diffusion models have recently achieved remarkable results in video generation. Despite their encouraging performance, most of these models are mainly designed and trained for short video generation, leading to challenges in maintaining temporal consistency and visual details in long video generation. In this paper, through theoretical analysis of the mechanisms behind video generation, we identify two key challenges that hinder short-to-long generalization, namely, temporal position ambiguity and information dilution. To address these challenges, we propose LongDiff, a novel training-free method that unlocks the potential of the off-the-shelf video diffusion models to achieve high-quality long video generation in one go. Extensive experiments demonstrate the efficacy of our method.
****

Feature Selection for Latent Factor Models
Rittwika Kansabanik, Adrian Barbu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30742-30751
Feature selection is crucial for pinpointing relevant features in high-dimensional datasets, mitigating the 'curse of dimensionality,' and enhancing machine learning performance. Traditional feature selection methods for classification use data from all classes to select features for each class.This paper explores feature selection methods that select features for each class separately, using class models based on low-rank generative methods and introducing a signal-to-noise ratio (SNR) feature selection criterion. This novel approach theoretically guarantees true feature recovery under certain assumptions and is shown to outperform

some existing feature selection methods on standard classification datasets.
****

Preserve or Modify? Context-Aware Evaluation for Balancing Preservation and Modification in Text-Guided Image Editing

Yoonjeon Kim, Soohyun Ryu, Yeonsung Jung, Hyunkoo Lee, Joowon Kim, June Yong Yang, Jaeryong Hwang, Eunho Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23474-23483

The development of vision-language and generative models has significantly advanced text-guided image editing, which seeks the preservation of core elements in the source image while implementing modifications based on the target text. However, existing metrics have a context-blindness problem, indiscriminately applying the same evaluation criteria on completely different pairs of source image and target text, biasing towards either modification or preservation. Directional CLIP similarity, the only metric that considers both source image and target text, is also biased towards modification aspects and attends to irrelevant editing regions of the image. We propose AugCLIP, a context-aware metric that adaptively coordinates preservation and modification aspects, depending on the specific context of a given source image and target text. This is done by deriving the CLIP representation of an ideally edited image, that preserves the source image with necessary modifications to align with target text. More specifically, using a multi-modal large language model, AugCLIP augments the textual descriptions of the source and target, then calculates a modification vector through a hyperplane that separates source and target attributes in CLIP space. Extensive experiments on five benchmark datasets, encompassing a diverse range of editing scenarios, show that AugCLIP aligns remarkably well with human evaluation standards, outperforming existing metrics. The code is available at https://github.com/augclip/augclip_eval.
****

Mask-Adapter: The Devil is in the Masks for Open-Vocabulary Segmentation

Yongkang Li, Tianheng Cheng, Bin Feng, Wenyu Liu, Xinggang Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14998-15008

Recent open-vocabulary segmentation methods adopt mask generators to predict segmentation masks and leverage pre-trained vision-language models, *e.g.*, CLIP, to classify these masks via mask pooling.Although these approaches show promising results, it is counterintuitive that accurate masks often fail to yield accurate classification results through pooling CLIP image embeddings within the mask regions.In this paper, we reveal the performance limitations of mask pooling and introduce **Mask-Adapter**, a simple yet effective method to address these challenges in open-vocabulary segmentation.Compared to directly using proposal masks, our proposed Mask-Adapter extracts *semantic activation maps* from proposal masks, providing richer contextual information and ensuring alignment between masks and CLIP.Additionally, we propose a *mask consistency loss* that encourages proposal masks with similar IoUs to obtain similar CLIP embeddings to enhance models' robustness to varying predicted masks.Mask-Adapter integrates seamlessly into open-vocabulary segmentation methods based on mask pooling in a plug-and-play manner, delivering more accurate classification results. Extensive experiments across several zero-shot benchmarks demonstrate significant performance gains for the proposed Mask-Adapter on several well-established methods.Notably, Mask-Adapter also extends effectively to SAM and achieves impressive results on several open-vocabulary segmentation datasets. Code and models are available at https://github.com/hustvl/MaskAdapter.
****

MPDrive: Improving Spatial Understanding with Marker-Based Prompt Learning for Autonomous Driving

Zhiyuan Zhang, Xiaofan Li, Zhihao Xu, Wenjie Peng, Zijian Zhou, Miaojing Shi, Shuangping Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12089-12099

Autonomous driving visual question answering (AD-VQA) aims to answer questions related to perception, prediction, and planning based on given driving scene imag

es, heavily relying on the model's spatial understanding capabilities. Prior wor ks typically express spatial information through textual representations of coor dinates, resulting in semantic gaps between visual coordinate representations an d textual descriptions. This oversight hinders the accurate transmission of spat ial information and increases the expressive burden. To address this, we propose a novel Marker-based Prompt learning framework (MPDrive), which represents spat ial coordinates by concise visual markers, ensuring linguistic expressive consis tency and enhancing the accuracy of both visual perception and spatial expressio n in AD-VQA. Specifically, we create marker images by employing a detection expe rt to overlay object regions with numerical labels, converting complex textual c oordinate generation into straightforward text-based visual marker predictions. Moreover, we fuse original and marker images as scene-level features and integra te them with detection priors to derive instance-level features. By combining th ese features, we construct dual-granularity visual prompts that stimulate the LL M's spatial perception capabilities. Extensive experiments on the DriveLM and CO DA-LM datasets show that MPDrive achieves state-of-the-art performance, particul arly in cases requiring sophisticated spatial understanding.

****

Improving the Transferability of Adversarial Attacks on Face Recognition with Di verse Parameters Augmentation

Fengfan Zhou, Bangjie Yin, Hefei Ling, Qianyu Zhou, Wenxuan Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3516-3 527

Face Recognition (FR) models are vulnerable to adversarial examples that subtly manipulate benign face images, underscoring the urgent need to improve the trans ferability of adversarial attacks in order to expose the blind spots of these sy stems. Existing adversarial attack methods often overlook the potential benefits of augmenting the surrogate model with diverse initializations, which limits th e transferability of the generated adversarial examples. To address this gap, we propose a novel method called Diverse Parameters Augmentation (DPA) attack meth od, which enhances surrogate models by incorporating diverse parameter initializ ations, resulting in a broader and more diverse set of surrogate models. Specifi cally, DPA consists of two key stages: Diverse Parameters Optimization (DPO) and Hard Model Aggregation (HMA). In the DPO stage, we initialize the parameters of the surrogate model using both pre-trained and random parameters. Subsequently, we save the models in the intermediate training process to obtain a diverse set of surrogate models. During the HMA stage, we enhance the feature maps of the d iversified surrogate models by incorporating beneficial perturbations, thereby f urther improving the transferability. Experimental results demonstrate that our proposed attack method can effectively enhance the transferability of the crafte d adversarial face examples.

****

Adapting to Observation Length of Trajectory Prediction via Contrastive Learning

Ruiqi Qiu, Jun Gong, Xinyu Zhang, Siqi Luo, Bowen Zhang, Yi Cen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1645-16 54

The ability to adapt to varying observation lengths is crucial for human traject ory prediction tasks, particularly in scenarios with limited observation lengths or missing data. Existing approaches mainly focus on introducing novel architec tures or additional structural components, which substantially increase model co mplexity and present challenges for integration into existing models. We argue t hat current network architectures are sufficiently sophisticated to handle the O bservation Length Shift problem, with the key challenge lying in improving featu re representation for trajectories with limited lengths. To tackle this issue, w e introduce a general and effective contrastive learning approach, called Contra stive Learning for Length Shift (CLLS). By incorporating contrastive learning du ring the training phase, our method encourages the model to extract length-invar iant features, thus mitigating the impact of observation length variations. Furt hermore, to better accommodate length adaptation tasks, we introduce a lightweig ht RNN network that, combined with CLLS, achieves state-of-the-art performance i

n both general prediction and observation length shift tasks. Experimental resul
ts demonstrate that our approach outperforms existing methods across multiple wi
dely-used trajectory prediction datasets.
****

Fine-Grained Image-Text Correspondence with Cost Aggregation for Open-Vocabulary
 Part Segmentation

Jiho Choi, Seonho Lee, Minhyun Lee, Seungho Lee, Hyunjung Shim; Proceedings of t
he Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9782-979
3

Open-Vocabulary Part Segmentation (OVPS) is an emerging field for recognizing fi
ne-grained parts in unseen categories. We identify two primary challenges in OVP
S: (1) the difficulty in aligning part-level image-text correspondence, and (2)
the lack of structural understanding in segmenting object parts. To address thes
e issues, we propose PartCATSeg, a novel framework that integrates object-aware
part-level cost aggregation, compositional loss, and structural guidance from DI
NO. Our approach employs a disentangled cost aggregation strategy that handles o
bject and part-level costs separately, enhancing the precision of part-level seg
mentation. We also introduce a compositional loss to better capture part-object
relationships, compensating for the limited part annotations. Additionally, stru
ctural guidance from DINO features improves boundary delineation and inter-part
understanding. Extensive experiments on Pascal-Part-116, ADE20K-Part-234, and Pa
rtImageNet datasets demonstrate that our method significantly outperforms state-
of-the-art approaches, setting a new baseline for robust generalization to unsee
n part categories.
****

NitroFusion: High-Fidelity Single-Step Diffusion through Dynamic Adversarial Tra
ining

Dar-Yen Chen, Hmrishav Bandyopadhyay, Kai Zou, Yi-Zhe Song; Proceedings of the C
omputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7654-7663

We introduce NitroFusion, a fundamentally different approach to single-step diff
usion that achieves high-quality generation through a dynamic adversarial framew
ork. While one-step methods offer dramatic speed advantages, they typically suff
er from quality degradation compared to their multi-step counterparts. Just as a
 panel of art critics provides comprehensive feedback by specializing in differe
nt aspects like composition, color, and technique, our approach maintains a larg
e pool of specialized discriminator heads that collectively guide the generation
 process. Each discriminator group develops expertise in specific quality aspect
s at different noise levels, providing diverse feedback that enables high-fideli
ty one-step generation. Our framework combines: (i) a dynamic discriminator pool
 with specialized discriminator groups to improve generation quality, (ii) strat
egic refresh mechanisms to prevent discriminator overfitting, and (iii) global-l
ocal discriminator heads for multi-scale quality assessment, and unconditional/c
onditional training for balanced generation. Additionally, our framework uniquel
y supports flexible deployment through bottom-up refinement, allowing users to d
ynamically choose between 1-4 denoising steps with the same model for direct qua
lity-speed trade-offs. Through comprehensive experiments, we demonstrate that Ni
troFusion significantly outperforms existing single-step methods across multiple
 evaluation metrics, particularly excelling in preserving fine details and globa
l consistency.
****

ByTheWay: Boost Your Text-to-Video Generation Model to Higher Quality in a Train
ing-free Way

Jiazi Bu, Pengyang Ling, Pan Zhang, Tong Wu, Xiaoyi Dong, Yuhang Zang, Yuhang Ca
o, Dahua Lin, Jiaqi Wang; Proceedings of the Computer Vision and Pattern Recogni
tion Conference (CVPR), 2025, pp. 12999-13008

The text-to-video (T2V) generation models, offering convenient visual creation,
have recently garnered increasing attention. Despite their substantial potential
, the generated videos may present artifacts, including structural implausibilit
y, temporal inconsistency, and a lack of motion, often resulting in near-static
video. In this work, we have identified a correlation between the disparity of t

emporal attention maps across different blocks and the occurrence of temporal in consistencies. Additionally, we have observed that the energy contained within t he temporal attention maps is directly related to the magnitude of motion amplit ude in the generated videos. Based on these observations, we present ByTheWay, a training-free method to improve the quality of text-to-video generation without introducing additional parameters, augmenting memory or sampling time. Specific ally, ByTheWay is composed of two principal components: 1) Temporal Self-Guidanc e improves the structural plausibility and temporal consistency of generated vid eos by reducing the disparity between the temporal attention maps across various decoder blocks. 2) Fourier-based Motion Enhancement enhances the magnitude and richness of motion by amplifying the energy of the map. Extensive experiments de monstrate that ByTheWay significantly improves the quality of text-to-video gene ration with negligible additional cost.

****

CMMLoc: Advancing Text-to-PointCloud Localization with Cauchy-Mixture-Model Base d Framework

Yanlong Xu, Haoxuan Qu, Jun Liu, Wenxiao Zhang, Xun Yang; Proceedings of the Com puter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6637-6647

The goal of point cloud localization based on linguistic description is to ident ify a 3D position using textual description in large urban environments, which h as potential applications in various fields, such as determining the location fo r vehicle pickup or goods delivery. Ideally, for a textual description and its c orresponding 3D location, the objects around the 3D location should be fully des cribed in the text description. However, in practical scenarios, e.g., vehicle p ickup, passengers usually describe only the part of the most significant and nea rby surroundings instead of the entire environment. In response to this partiall y relevant challenge, we propose CMMLoc, an uncertainty-aware Cauchy-Mixture-Mod el (CMM) based framework for text-to-point-cloud Localization. To model the unce rtain semantic relations between text and point cloud, we integrate CMM constrai nts as a prior during the interaction between the two modalities. We further des ign a spatial consolidation scheme to enable adaptive aggregation of different 3 D objects with varying receptive fields. To achieve precise localization, we pro pose a cardinal direction integration module alongside a modality pre-alignment strategy, helping capture the spatial relationships among objects and bringing t he 3D objects closer to the text modality. Comprehensive experiments validate th at CMMLoc outperforms existing methods, achieving state-of-the-art results on th e KITTI360Pose dataset. Codes are available in this anonymous GitHub repository https://github.com/anonymous0819/CMMLoc.

****

Masked Point-Entity Contrast for Open-Vocabulary 3D Scene Understanding

Yan Wang, Baoxiong Jia, Ziyu Zhu, Siyuan Huang; Proceedings of the Computer Visi on and Pattern Recognition Conference (CVPR), 2025, pp. 14125-14136

Open-vocabulary 3D scene understanding is pivotal for enhancing physical intelli gence, as it enables embodied agents to interpret and interact dynamically withi n real-world environments. This paper introduces MPEC, a novel Masked Point-Enti ty Contrastive learning method for open-vocabulary 3D semantic segmentation that leverages both 3D entity-language alignment and point-entity consistency across different point cloud views to foster entity-specific feature representations. Our method improves semantic discrimination and enhances the differentiation of unique instances, achieving state-of-the-art results on ScanNet for open-vocabul ary 3D semantic segmentation and demonstrating superior zero-shot scene understa nding capabilities. Extensive fine-tuning experiments on 8 datasets, spanning fr om low-level perception to high-level reasoning tasks, showcase the potential of learned 3D features, driving consistent performance gains across varied 3D scen e understanding tasks.

****

Decoupling Training-Free Guided Diffusion by ADMM

Youyuan Zhang, Zehua Liu, Zenan Li, Zhaoyu Li, James J. Clark, Xujie Si; Proceed ings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23292-23302

In this paper, we consider the conditional generation problem by guiding off-the-shelf unconditional diffusion models with differentiable loss functions in a plug-and-play fashion. While previous research has primarily focused on balancing the unconditional diffusion model and the guided loss through a tuned weight hyperparameter, we propose a novel framework that distinctly decouples these two components. Specifically, we introduce two variables x and z, to represent the generated samples governed by the unconditional generation model and the guidance function, respectively. This decoupling reformulates conditional generation into two manageable subproblems, unified by the constraint $x = z$. Leveraging this setup, we develop a new algorithm based on the Alternating Direction Method of Multipliers (ADMM) to adaptively balance these components. Additionally, we establish the equivalence between the diffusion reverse step and the proximal operator of ADMM and provide a detailed convergence analysis of our algorithm under certain mild assumptions. Our experiments demonstrate that our proposed method \OurMethod consistently generates high-quality samples while ensuring strong adherence to the conditioning criteria. It outperforms existing methods across a range of conditional generation tasks, including image generation with various guidance and controllable motion synthesis.
****

On the Generalization of Handwritten Text Recognition Models
Carlos Garrido-Munoz, Jorge Calvo-Zaragoza; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15275-15286
Recent advances in Handwritten Text Recognition (HTR) have led to significant reductions in transcription errors on standard benchmarks under the i.i.d. assumption, thus focusing on minimizing in-distribution (ID) errors.However, this assumption does not hold in real-world applications, which has motivated HTR research to explore Transfer Learning and Domain Adaptation techniques. In this work, we investigate the unaddressed limitations of HTR models in generalizing to out-of-distribution (OOD) data. We adopt the challenging setting of Domain Generalization, where models are expected to generalize to OOD data without any prior access. To this end, we analyze 336 OOD cases from eight state-of-the-art HTR models across seven widely used datasets, spanning five languages. Additionally, we study how HTR models leverage synthetic data to generalize. We reveal that the most significant factor for generalization lies in the textual divergence between domains, followed by visual divergence. We demonstrate that the error of HTR models in OOD scenarios can be reliably estimated, with discrepancies falling below 10 points in 70% of cases. We identify the underlying limitations of HTR models, laying the foundation for future research to address this challenge.
****

SwiftEdit: Lightning Fast Text-Guided Image Editing via One-Step Diffusion
Trong-Tung Nguyen, Quang Nguyen, Khoi Nguyen, Anh Tran, Cuong Pham; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21492-21501
Recent advances in text-guided image editing enable users to perform image edits through simple text inputs, leveraging the extensive priors of multi-step diffusion-based text-to-image models. However, these methods often fall short of the speed demands required for real-world and on-device applications due to the costly multi-step inversion and sampling process involved. In response to this, we introduce SwiftEdit, a simple yet highly efficient editing tool that achieve instant text-guided image editing in 0.23s. The advancement of SwiftEdit lies in its two novel contributions: a one-step inversion framework that enables one-step image reconstruction via inversion and a mask-guided editing technique with our proposed attention rescaling mechanism to perform localized image editing. Extensive experiments are provided to demonstrate the effectiveness and efficiency of SwiftEdit. In particular, SwiftEdit enables instant text-guided image editing, which is extremely faster than previous multi-step methods (at least 50 times faster) while maintain a competitive performance in editing results. Our project is at https://swift-edit.github.io/.
****

Learning from Synchronization: Self-Supervised Uncalibrated Multi-View Person As

sociation in Challenging Scenes
Keqi Chen, Vinkle Srivastav, Didier Mutter, Nicolas Padoy; Proceedings of the Co mputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24419-24428
Multi-view person association is a fundamental step towards multi-view analysis of human activities. Although the person re-identification features have been pr oven effective, they become unreliable in challenging scenes where persons share similar appearances. Therefore, cross-view geometric constraints are required f or a more robust association. However, most existing approaches are either fully -supervised using ground-truth identity labels or require calibrated camera para meters that are hard to obtain. In this work, we investigate the potential of le arning from synchronization, and propose a self-supervised uncalibrated multi-vi ew person association approach, Self-MVA, without using any annotations. Specifi cally, we propose a self-supervised learning framework, consisting of an encoder -decoder model and a self-supervised pretext task, cross-view image synchronizat ion, which aims to distinguish whether two images from different views are captu red at the same time. The model encodes each person's unified geometric and appe arance features, and we train it by utilizing synchronization labels for supervi sion after applying Hungarian matching to bridge the gap between instance-wise a nd image-wise distances. To further reduce the solution space, we propose two t ypes of self-supervised linear constraints: multi-view re-projection and pairwis e edge association. Extensive experiments on three challenging public benchmark datasets (WILDTRACK, MVOR, and SOLDIERS) show that our approach achieves state-o f-the-art results, surpassing existing unsupervised and fully-supervised approac hes. Code is available at https://github.com/CAMMA-public/Self-MVA.
****

RC-AutoCalib: An End-to-End Radar-Camera Automatic Calibration Network
Van-Tin Luu, Yon-Lin Cai, Vu-Hoang Tran, Wei-Chen Chiu, Yi-Ting Chen, Ching-Chun Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (C VPR), 2025, pp. 6700-6709
This paper presents a groundbreaking approach - the first online automatic geome tric calibration method for radar and camera systems. Given the significant data sparsity and measurement uncertainty in radar height data, achieving automatic calibration during system operation has long been a challenge. To address the sp arsity issue, we propose a Dual-Perspective representation that gathers features from both frontal and bird's-eye views. The frontal view contains rich but sens itive height information, whereas the bird's-eye view provides robust features a gainst height uncertainty. We thereby propose a novel Selective Fusion Mechanism to identify and fuse reliable features from both perspectives, reducing the eff ect of height uncertainty. Moreover, for each view, we incorporate a Multi-Modal Cross-Attention Mechanism to explicitly find location correspondences through c ross-modal matching. During the training phase, we also design a Noise-Resistant Matcher to provide better supervision and enhance the robustness of the matchin g mechanism against sparsity and height uncertainty. Our experimental results, t ested on the nuScenes dataset, demonstrate that our method significantly outperf orms previous radar-camera auto-calibration methods, as well as existing state-o f-the-art LiDAR-camera calibration techniques, establishing a new benchmark for future research.
****

Argus: A Compact and Versatile Foundation Model for Vision
Weiming Zhuang, Chen Chen, Zhizhong Li, Sina Sajadmanesh, Jingtao Li, Jiabo Huan g, Vikash Sehwag, Vivek Sharma, Hirotaka Shinozaki, Felan Carlo Garcia, Yihao Zh an, Naohiro Adachi, Ryoji Eki, Michael Spranger, Peter Stone, Lingjuan Lyu; Proc eedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4418-4429
While existing vision and multi-modal foundation models can handle multiple comp uter vision tasks, they often suffer from significant limitations, including hug e demand for data and computational resources during training and inconsistent p erformance across vision tasks at deployment time. To address these challenges, we introduce Argus (The name comes from Argus Panoptes -- a hundred-eyed giant w ith "all-seeing" capability in Greek mythology), a compact and versatile vision

foundation model designed to support a wide range of vision tasks through a unified multitask architecture. Argus employs a two-stage training strategy: (i) multitask pretraining over core vision tasks with a shared backbone that includes a lightweight adapter to inject task-specific inductive biases, and (ii) scalable and efficient adaptation to new tasks by fine-tuning only the task-specific decoders. Extensive evaluations demonstrate that Argus, despite its relatively compact and training-efficient design of merely 100M backbone parameters (only 13.6% of which are trained using 1.6M images), competes with and even surpasses much larger models. Compared to state-of-the-art foundation models, Argus not only covers a broader set of vision tasks but also matches or outperforms the models with similar sizes on 12 tasks. We expect that Argus will accelerate the real-world adoption of vision foundation models in resource-constrained scenarios.
****

CLIP-driven Coarse-to-fine Semantic Guidance for Fine-grained Open-set Semi-supervised Learning
Xiaokun Li, Yaping Huang, Qingji Guan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30312-30321
Fine-grained open-set semi-supervised learning (OSSL) investigates a practical scenario where unlabeled data may contain fine-grained out-of-distribution (OOD) samples. Due to the subtle visual differences among in-distribution (ID) samples, as well as between ID and OOD samples, it is extremely challenging to separate the ID and OOD samples. Due to the subtle visual differences among in-distribution (ID) and OOD samples. Recent Vision-Language Models, such as CLIP, have shown excellent generalization capabilities. However, it tends to focus on general attributes, and thus is insufficient to distinguish the fine-grained details. To tackle the issues, in this paper, we propose a novel CLIP-driven coarse-to-fine semantic-guided framework, named CFSG-CLIP, to progressively focus on the distinctive fine-grained clues. Specifically, CFSG-CLIP comprises a coarse-guidance branch and a fine-guidance branch derived from the pre-trained CLIP model. In the coarse-guidance branch, we design a semantic filtering module to initially filter and highlight local visual features guided by cross-modality features. Then, in the fine-guidance branch, we further design a visual-semantic injection strategy, which embeds category-related visual cues into the visual encoder to further refine the local visual features. By the designed dual-guidance framework, local subtle cues are progressively discovered to distinct the subtle difference between ID and OOD samples. Extensive experiments demonstrate that CFSG-CLIP achieves competitive performance on multiple fine-grained datasets. The source code is available at https://github.com/LxxxxK/CFSG-CLIP.
****

InsTaG: Learning Personalized 3D Talking Head from Few-Second Video
Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Jun Zhou, Lin Gu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10690-10700
Despite exhibiting impressive performance in synthesizing lifelike personalized 3D talking heads, prevailing methods based on radiance fields suffer from high demands for training data and time for each new identity. This paper introduces InsTaG, a 3D talking head synthesis framework that allows a fast learning of realistic personalized 3D talking head from few training data. Built upon a lightweight 3DGS person-specific synthesizer with universal motion priors, InsTaG achieves high-quality and fast adaptation while preserving high-level personalization and efficiency. As preparation, we first propose an Identity-Free Pre-training strategy that enables the pre-training of the person-specific model and encourages the collection of universal motion priors from long-video data corpus. To fully exploit the universal motion priors to learn an unseen new identity, we then present a Motion-Aligned Adaptation strategy to adaptively align the target head to the pre-trained field, and constrain a robust dynamic head structure under few training data. Experiments demonstrate our outstanding performance and efficiency under various data scenarios to render high-quality personalized talking heads. Project page: https://fictionarry.github.io/InsTaG/.
****

Sampling Innovation-Based Adaptive Compressive Sensing
Zhifu Tian, Tao Hu, Chaoyang Niu, Di Wu, Shu Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2387-2397
Scene-aware Adaptive Compressive Sensing (ACS) has attracted significant interest due to its promising capability for efficient and high-fidelity acquisition of scene images. ACS typically prescribes adaptive sampling allocation (ASA) based on previous samples in the absence of ground truth. However, when confronting unknown scenes, existing ACS methods often lack accurate judgment and robust feedback mechanisms for ASA, thus limiting the high-fidelity sensing of the scene. In this paper, we introduce a Sampling Innovation-Based ACS (SIB-ACS) method that can effectively identify and allocate sampling to challenging image reconstruction areas, culminating in high-fidelity image reconstruction. An innovation criterion is proposed to judge ASA by predicting the decrease in image reconstruction error attributable to sampling increments, thereby directing more samples towards regions where the reconstruction error diminishes significantly. A sampling innovation-guided multi-stage adaptive sampling (AS) framework is proposed, which iteratively refines the ASA through a multi-stage feedback process. For image reconstruction, we propose a Principal Component Compressed Domain Network (PCCD-Net), which efficiently and faithfully reconstructs images under AS scenarios. Extensive experiments demonstrate that the proposed SIB-ACS method significantly outperforms the state-of-the-art methods in terms of image reconstruction fidelity and visual effects. Codes are available at https://github.com/giant-pandada/SIB-ACS_CVPR2025.
****

A Simple Data Augmentation for Feature Distribution Skewed Federated Learning
Yunlu Yan, Huazhu Fu, Yuexiang Li, Jinheng Xie, Jun Ma, Guang Yang, Lei Zhu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25749-25758
Federated Learning (FL) facilitates collaborative learning among multiple clients in a distributed manner and ensures the security of privacy. However, its performance inevitably degrades with non-Independent and Identically Distributed (non-IID) data. In this paper, we focus on the feature distribution skewed FL scenario, a common non-IID situation in real-world applications where data from different clients exhibit varying underlying distributions. This variation leads to feature shift, which is a key issue of this scenario. While previous works have made notable progress, few pay attention to the data itself, i.e., the root of this issue. The primary goal of this paper is to mitigate feature shift from the perspective of data. To this end, we propose a simple yet remarkably effective input-level data augmentation method, namely FedRDN, which randomly injects the statistical information of the local distribution from the entire federation into the client's data. This is beneficial to improve the generalization of local feature representations, thereby mitigating feature shift. Moreover, our FedRDN is a plug-and-play component, which can be seamlessly integrated into the data augmentation flow with only a few lines of code. Extensive experiments on several datasets show that the performance of various representative FL methods can be further improved by integrating our FedRDN, demonstrating its effectiveness, strong compatibility and generalizability. Code is available at https://github.com/IAMJackYan/FedRDN.
****

MotionBench: Benchmarking and Improving Fine-grained Video Motion Understanding for Vision Language Models
Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, Jie Tang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8450-8460
In the quest for artificial general intelligence, Multi-modal Large Language Models (MLLMs) have emerged as a focal point in recent advancements. However, the predominant focus remains on developing their capabilities in static image understanding. The potential of MLLMs in processing sequential visual data is still insufficiently explored, highlighting the absence of a comprehensive, high-quality assessment of their performance. In this paper, we introduce Video-MME, the fir

st-ever full-spectrum, Multi-Modal Evaluation benchmark of MLLMs in Video analysis. Our work distinguishes from existing benchmarks through four key features: 1) Diversity in video types, spanning 6 primary visual domains with 30 subfields to ensure broad scenario generalizability; 2) Duration in temporal dimension, encompassing both short-, medium-, and long-term videos, ranging from 11 seconds to 1 hour, for robust contextual dynamics; 3) Breadth in data modalities, integrating multi-modal inputs besides video frames, including sub- titles and audios, to unveil the all-round capabilities of MLLMs; 4) Quality in annotations, utilizing rigorous manual labeling by expert annotators to facilitate precise and reliable model assessment. 900 videos with a total of 254 hours are manually selected and annotated by repeatedly viewing all the video content, resulting in 2,700 question-answer pairs. With Video-MME, we extensively evaluate various state-of-the-art MLLMs, including GPT-4 series and Gemini 1.5 Pro, as well as open-source image models like InternVL-Chat-V1.5 and video models like LLaVA-NeXT-Video. Our experiments reveal that Gemini 1.5 Pro is the best-performing commercial model, significantly outperforming the open-source models with an average accuracy of 75%, compared to 71.9% for GPT-4o. The results also demonstrate that Video-MME is a universal benchmark, which applies to both image and video MLLMs. Further analysis indicates that subtitle and audio information could significantly enhance video understanding. Besides, a decline in MLLM performance is observed as video duration increases for all models. Our dataset along with these findings underscores the need for further improvements in handling longer sequences and multi-modal data, shedding light on future MLLM development. Project page: https://video-mme.github.io.

****

Benchmarking Large Vision-Language Models via Directed Scene Graph for Comprehensive Image Captioning

Fan Lu, Wei Wu, Kecheng Zheng, Shuailei Ma, Biao Gong, Jiawei Liu, Wei Zhai, Yang Cao, Yujun Shen, Zheng-Jun Zha; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19618-19627

Generating detailed captions comprehending text-rich visual content in images has received growing attention for Large Vision-Language Models (LVLMs). However, few studies have developed benchmarks specifically tailored for detailed captions to measure their accuracy and comprehensiveness. In this paper, we introduce a detailed caption benchmark, termed as CompreCap, to evaluate the visual context from a directed scene graph view. Concretely, we first manually segment the image into semantically meaningful regions (i.e., semantic segmentation mask) according to common-object vocabulary, while also distinguishing attributes of objects within all those regions. Then directional relation labels of these objects are annotated to compose a directed scene graph that can well encode rich compositional information of the image. Based on our directed scene graph, we develop a pipeline to assess the generated detailed captions from LVLMs on multiple levels, including the object-level coverage, the accuracy of attribute descriptions, the score of key relationships, etc. Experimental results on the CompreCap dataset confirm that our evaluation method aligns closely with human evaluation scores across LVLMs. We will release the code and the dataset to support the community.

****

ART: Anonymous Region Transformer for Variable Multi-Layer Transparent Image Generation

Yifan Pu, Yiming Zhao, Zhicong Tang, Ruihong Yin, Haoxing Ye, Yuhui Yuan, Dong Chen, Jianmin Bao, Sirui Zhang, Yanbin Wang, Lin Liang, Lijuan Wang, Ji Li, Xiu Li, Zhouhui Lian, Gao Huang, Baining Guo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7952-7962

Multi-layer image generation is a fundamental task that enables users to isolate, select, and edit specific image layers, thereby revolutionizing interactions with generative models. In this paper, we introduce the Anonymous Region Transformer (ART), which facilitates the direct generation of variable multi-layer transparent images based on a global text prompt and an anonymous region layout. Inspired by Schema theory, this anonymous region layout allows the generative model

to autonomously determine which set of visual tokens should align with which tex
t tokens, which is in contrast to the previously dominant semantic layout for th
e image generation task. In addition, the layer-wise region crop mechanism, whic
h only selects the visual tokens belonging to each anonymous region, significant
ly reduces attention computation costs and enables the efficient generation of i
mages with numerous distinct layers (e.g., 50+). When compared to the full atten
tion approach, our method is over 12 times faster and exhibits fewer layer confl
icts. Furthermore, we propose a high-quality multi-layer transparent image autoe
ncoder that supports the direct encoding and decoding of the transparency of var
iable multi-layer images in a joint manner. By enabling precise control and scal
able layer generation, ART establishes a new paradigm for interactive content cr
eation.
****

Rotation-Equivariant Self-Supervised Method in Image Denoising
Hanze Liu, Jiahong Fu, Qi Xie, Deyu Meng; Proceedings of the Computer Vision and
 Pattern Recognition Conference (CVPR), 2025, pp. 12720-12730
Self-supervised image denoising methods have garnered significant research atten
tion in recent years, for this kind of method reduces the requirement of large t
raining datasets.Compared to supervised methods, self-supervised methods rely mo
re on the prior embedded in deep networks themselves. As a result, most of the s
elf-supervised methods are designed with Convolution Neural Networks(CNNs) archi
tectures, which well capture one of the most important image prior, translation
equivariant prior. Inspired by the great success achieved by the introduction of
 translational equivariance, in this paper, we explore the way to further incorp
orate another important image prior. Specifically, we first apply high-accuracy
rotation equivariant convolution to self-supervised image denoising. Through rig
orous theoretical analysis, we have proved that simply replacing all the convolu
tion layers with rotation equivariant convolution layers would modify the networ
k into its rotation equivariant version.To the best of our knowledge, this is th
e first time that rotation equivariant image prior is introduced to self-supervi
sed image denoising at the network architecture level with a comprehensive theor
etical analysis of equivariance errors, whichoffers a new perspective to the fie
ld of self-supervised image denoising.Moreover, to further improve the performan
ce, we design a new mask mechanism to fusion the output of rotation equivariant
network and vanilla CNN-based network, and construct an adaptive rotation equiva
riant framework. Through extensive experiments on three typical methods, we have
 demonstrated the effectiveness of the proposed method.
****

ArcPro: Architectural Programs for Structured 3D Abstraction of Sparse Points
Qirui Huang, Runze Zhang, Kangjun Liu, Minglun Gong, Hao Zhang, Hui Huang; Proce
edings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, p
p. 6563-6572
We introduce ArcPro, a novel learning framework built on architectural programs
to recover structured 3D abstractions from highly sparse and low-quality point c
louds. Specifically, we design a domain-specific language (DSL) to hierarchicall
y represent building structures as a program, which can be efficiently converted
 into a mesh. We bridge feedforward and inverse procedural modeling by using a f
eedforward process for training data synthesis, allowing the network to make rev
erse predictions. We train an encoder-decoder on the points-program pairs to est
ablish a mapping from unstructured point clouds to architectural programs, where
 a 3D convolutional encoder extracts point cloud features and a transformer deco
der autoregressively predicts the programs in a tokenized form. Inference by our
 method is highly efficient and produces plausible and faithful 3D abstractions.
 Comprehensive experiments demonstrate that ArcPro outperforms both traditional
architectural proxy reconstruction and learning-based abstraction methods. We fu
rther explore its potential when working with multi-view image and natural langu
age inputs.
****

GLane3D: Detecting Lanes with Graph of 3D Keypoints
Halil ■brahim Öztürk, Muhammet Esat Kalfao■lu, Ozsel Kilinc; Proceedings of the

Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27508-27518
Accurate and efficient lane detection in 3D space is essential for autonomous driving systems, where robust generalization is the foremost requirement for 3D lane detection algorithms. Considering the extensive variation in lane structures worldwide, achieving high generalization capacity is particularly challenging, as algorithms must accurately identify a wide variety of lane patterns worldwide. Traditional top-down approaches rely heavily on learning lane characteristics from training datasets, often struggling with lanes exhibiting previously unseen attributes. To address this generalization limitation, we propose a method that detects keypoints of lanes and subsequently predicts sequential connections between them to construct complete 3D lanes. Each key point is essential for maintaining lane continuity, and we predict multiple proposals per keypoint by allowing adjacent grids to predict the same keypoint using an offset mechanism. PointNMS is employed to eliminate overlapping proposal keypoints, reducing redundancy in the estimated BEV graph and minimizing computational overhead from connection estimations. Our model surpasses previous state-of-the-art methods on both the Apollo and OpenLane datasets, demonstrating superior F1 scores and a strong generalization capacity when models trained on OpenLane are evaluated on the Apollo dataset, compared to prior approaches.
****

Minimal Interaction Seperated Tuning: A New Paradigm for Visual Adaptation
Ningyuan Tang, Minghao Fu, Jianxin Wu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25208-25217
The rapid scaling of large vision pretrained models makes fine-tuning tasks more and more difficult on devices with low computational resources. We explore a new visual adaptation paradigm called separated tuning, which treats large pretrained models as standalone feature extractors that run on powerful cloud servers. The fine-tuning carries out on devices which possess only low computational resources (slow CPU, no GPU, small memory, etc.) Existing methods that are potentially suitable for our separated tuning paradigm are discussed. But, three major drawbacks hinder their application in separated tuning: low adaptation capability, large adapter network, and in particular, high information transfer overhead. To address these issues, we propose Minimal Interaction Separated Tuning, or MIST, which reveals that the sum of intermediate features from pretrained models not only has minimal information transfer but also has high adaptation capability. With a lightweight attention-based adaptor network, MIST achieves information transfer efficiency, parameter efficiency, computational and memory efficiency, and at the same time demonstrates competitive results on various visual adaptation benchmarks.
****

Hardware-Rasterized Ray-Based Gaussian Splatting
Samuel Rota Bulò, Nemanja Bartolovic, Lorenzo Porzi, Peter Kontschieder; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 485-494
We present a novel, hardware-rasterized rendering approach for ray-based 3D Gaussian Splatting (RayGS), obtaining both fast and high-quality results for novel view synthesis. Our work contains a mathematically rigorous and geometrically intuitive derivation about how to efficiently estimate all relevant quantities for rendering RayGS models, structured with respect to standard hardware rasterization shaders. Our solution is the first enabling rendering RayGS models at sufficiently high frame rates to support quality-sensitive applications like Virtual and Mixed Reality. Our second contribution enables alias-free rendering for RayGS, by addressing MIP-related issues arising when rendering diverging scales during training and testing. We demonstrate significant performance gains, across different benchmark scenes, while retaining state-of-the-art appearance quality of RayGS.
****

FlashSloth : Lightning Multimodal Large Language Models via Embedded Visual Compression
Bo Tong, Bokai Lai, Yiyi Zhou, Gen Luo, Yunhang Shen, Ke Li, Xiaoshuai Sun, Rong

rong Ji; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14570-14581
Despite a big leap forward in capability, multimodal large language models (MLLM s) tend to behave like a sloth in practical use, i.e., slow response and large l atency. Recent efforts are devoted to building tiny MLLMs for better efficiency, but the plethora of visual tokens still used limit their actual speedup. In thi s paper, we propose a powerful and fast tiny MLLM called FlashSloth. Different f rom previous efforts, FlashSloth focuses on improving the descriptive power of v isual tokens in the process of compressing their redundant semantics. In particu lar, FlashSloth introduces embedded visual compression designs to capture both v isually salient and instruction-related image information, so as to achieving su perior multimodal performance with fewer visual tokens. Extensive experiments ar e conducted to validate the proposed FlashSloth, and a bunch of tiny but strong MLLMs are also comprehensively compared, e.g., InternVL-2, MiniCPM-V2 and Qwen2- VL. The experimental results show that compared with these advanced tiny MLLMs, our FlashSloth can greatly reduce the number of visual tokens, training memory a nd computation complexity while retaining high performance on various VL tasks. Our code is released at: https://github.com/codefanw/FlashSloth.
****

FreqDebias: Towards Generalizable Deepfake Detection via Consistency-Driven Freq uency Debiasing
Hossein Kashiani, Niloufar Alipour Talemi, Fatemeh Afghah; Proceedings of the Co mputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8775-8785
Deepfake detectors often struggle to generalize to novel forgery types due to bi ases learned from limited training data. In this paper, we identify a new type o f model bias in the frequency domain, termed spectral bias, where detectors over ly rely on specific frequency bands, restricting their ability to generalize acr oss unseen forgeries. To address this, we propose FreqDebias, a frequency debias ing framework that mitigates spectral bias through two complementary strategies. First, we introduce a novel Forgery Mixup (Fo-Mixup) augmentation, which dynami cally diversifies frequency characteristics of training samples. Second, we inco rporate a dual consistency regularization (CR), which enforces both local consis tency using class activation maps (CAMs) and global consistency through a von Mi ses-Fisher (vMF) distribution on a hyperspherical embedding space. This dual CR mitigates over-reliance on certain frequency components by promoting consistent representation learning under both local and global supervision. Extensive exper iments show that FreqDebias significantly enhances cross-domain generalization a nd outperforms state-of-the-art methods in both cross-domain and in-domain setti ngs.
****

Multi-subject Open-set Personalization in Video Generation
Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Yuwei Fang, Kwot Sin Lee, Ivan Skorokhodov, Kfir Aberman, Jun-Yan Zhu, Ming-Hsuan Yang, Sergey Tulyakov; P roceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 202 5, pp. 6099-6110
Video personalization methods allow us to synthesize videos with specific concep ts such as people, pets, and places. However, existing methods often focus on li mited domains, require time-consuming optimization per subject, or support only a single subject. We present Video Alchemist--a video model with built-in multi- subject, open-set personalization capabilities for both foreground objects and b ackground, eliminating the need for time-consuming test-time optimization. Our m odel is built on a new Diffusion Transformer module that fuses each conditional reference image and its corresponding subject-level text prompt with cross-atten tion layers. Developing such a large model presents two main challenges: dataset and evaluation. First, as paired datasets of reference images and videos are ex tremely hard to collect, we sample selected video frames as reference images and synthesize a clip of the target video. However, while models can easily denoise training videos given reference frames, they fail to generalize to new contexts . To mitigate this issue, we design a new automatic data construction pipeline w ith extensive image augmentations. Second, evaluating open-set video personaliza

tion is a challenge in itself. To address this, we introduce a personalization b enchmark that focuses on accurate subject fidelity and supports diverse personal ization scenarios. Finally, our extensive experiments show that our method signi ficantly outperforms existing personalization methods in both quantitative and q ualitative evaluations.
****

Wav2Sem: Plug-and-Play Audio Semantic Decoupling for 3D Speech-Driven Facial Ani mation
Hao Li, Ju Dai, Xin Zhao, Feng Zhou, Junjun Pan, Lei Li; Proceedings of the Comp uter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 183-192
In 3D speech-driven facial animation generation, existing methods commonly emplo y pre-trained self-supervised audio models as encoders. However, due to the prev alence of phonetically similar syllables with distinct lip shapes in language, t hese near-homophone syllables tend to exhibit significant coupling in self-super vised audio feature spaces, leading to the averaging effect in subsequent lip mo tion generation. To address this issue, this paper proposes a plug-and-play sema ntic decorrelation module--Wav2Sem. This module extracts semantic features corre sponding to the entire audio sequence, leveraging the added semantic information to decorrelate audio encodings within the feature space, thereby achieving more expressive audio features. Extensive experiments across multiple Speech-driven models indicate that the Wav2Sem module effectively decouples audio features, si gnificantly alleviating the averaging effect of phonetically similar syllables i n lip shape generation, thereby enhancing the precision and naturalness of facia l animations. Our source code is available at https://github.com/wslh852/Wav2Sem .git.
****

Attraction Diminishing and Distributing for Few-Shot Class-Incremental Learning
Li-Jun Zhao, Zhen-Duo Chen, Yongxin Wang, Xin Luo, Xin-Shun Xu; Proceedings of t he Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25657-25 666
Few-Shot Class-Incremental Learning (FSCIL) aims to continuously learn novel cla sses with limited samples after pre-training on a set of base classes. To avoid catastrophic forgetting and overfitting, most FSCIL methods first train the mode l on the base classes and then freeze the feature extractor in the incremental s essions. However, the reliance on nearest neighbor classification makes FSCIL pr one to the hubness phenomenon, which negatively impacts performance in this dyna mic and open scenario. While recent methods attempt to adapt to the dynamic and open nature of FSCIL, they are often limited to biased optimizations to the feat ure space. In this paper, we pioneer the theoretical analysis of the inherent hu bness in FSCIL. To mitigate the negative effects of hubness, we propose a novel Attraction Diminishing and Distributing (D2A) method from the essential perspect ives of distance metric and feature space. Extensive experimental results demons trate that our method can broadly and significantly improve the performance of e xisting methods.
****

4DTAM: Non-Rigid Tracking and Mapping via Dynamic Surface Gaussians
Hidenobu Matsuki, Gwangbin Bae, Andrew J. Davison; Proceedings of the Computer V ision and Pattern Recognition Conference (CVPR), 2025, pp. 26921-26932
We propose the first 4D tracking and mapping method that jointly performs camera localization and non-rigid surface reconstruction via differentiable rendering. Our approach captures 4D scenes from an online stream of color images with dept h measurements or predictions by simultaneously optimizing scene geometry, appea rance, dynamics, and camera ego-motion. Although natural environments exhibit co mplex non-rigid motions, 4D-SLAM remains relatively underexplored due to its inh erent challenges; even with 2.5D signals, the problem is ill-posed because of th e high dimensionality of the optimization space. To overcome these challenges, w e first introduce a SLAM method based on Gaussian surface primitives that levera ges depth signals more effectively than 3D Gaussians, thereby achieving accurate surface reconstruction. To further model non-rigid deformations, we employ a wa rp-field represented by a multi-layer perceptron (MLP) and introduce a novel cam

era pose estimation technique along with surface regularization terms that facil itate spatio-temporal reconstruction. In addition to these algorithmic challenge s, a significant hurdle in 4D-SLAM research is the lack of publicly available da tasets with reliable ground truth and evaluation protocols. To address this, we present a novel synthetic dataset of everyday objects with diverse motions, leve raging large-scale object models and animation modeling. In summary, we open up the modern 4D-SLAM research by introducing a novel method and evaluation protoco ls grounded in modern vision and rendering techniques.
****

T2SG: Traffic Topology Scene Graph for Topology Reasoning in Autonomous Driving
Changsheng Lv, Mengshi Qi, Liang Liu, Huadong Ma; Proceedings of the Computer Vi sion and Pattern Recognition Conference (CVPR), 2025, pp. 17197-17206
Understanding the traffic scenes and then generating high-definition (HD) maps p resent significant challenges in autonomous driving. In this paper, we defined a novel \underline T raffic \underline T opology \underline S cene \underline G raph (\text T ^2\text SG), a unified scene graph explicitly modeling the lane, controlled and guided by different road signals (e.g. , right turn), and topol ogy relationships among them, which is always ignored by previous high-definitio n (HD) mapping methods. For the generation of \text T ^2\text SG , we propose To poFormer, a novel one-stage \underline Topo logy Scene Graph Trans\underline Fo rmer with two newly-designed layers. Specifically, TopoFormer incorporates a L ane Aggregation Layer (LAL) that leverages the geometric distance among the cent erline of lanes to guide the aggregation of global information. Furthermore, we proposed a Counterfactual Intervention Layer (CIL) to model the reasonable road structure (e.g. , intersection, straight) among lanes under counterfactual inte rvention. Then the generated \text T ^2\text SG can provide a more accurate and explainable description of the topological structure in traffic scenes. Experim ental results demonstrate that TopoFormer outperforms existing methods on the \t ext T ^2\text SG generation task, and the generated \text T ^2\text SG signifi cantly enhances traffic topology reasoning in downstream tasks, achieving a stat e-of-the-art performance of 46.3 OLS on the OpenLane-V2 benchmark. Our source co de is available at https://github.com/MICLAB-BUPT/T2SG.
****

Unseen Visual Anomaly Generation
Han Sun, Yunkang Cao, Hao Dong, Olga Fink; Proceedings of the Computer Vision an d Pattern Recognition Conference (CVPR), 2025, pp. 25508-25517
Visual anomaly detection (AD) presents significant challenges due to the scarcit y of anomalous data samples. While numerous works have been proposed to synthesi ze anomalous samples, these synthetic anomalies often lack authenticity or requi re extensive training data, limiting their applicability in real-world scenarios . In this work, we propose Anomaly Anything (AnomalyAny), a novel framework that leverages Stable Diffusion (SD)'s image generation capabilities to generate div erse and realistic unseen anomalies. By conditioning on a single normal sample d uring test time, AnomalyAny is able to generate unseen anomalies for arbitrary o bject types with text descriptions. Within AnomalyAny, we propose attention-guid ed anomaly optimization to direct SD's attention on generating hard anomaly conc epts. Additionally, we introduce prompt-guided anomaly refinement, incorporating detailed descriptions to further improve the generation quality. Extensive expe riments on MVTec AD and VisA datasets demonstrate AnomalyAny's ability in genera ting high-quality unseen anomalies and its effectiveness in enhancing downstream AD performance. Our demo and code are available at https://hansunhayden.github. io/CUT.github.io.
****

T2ICount: Enhancing Cross-modal Understanding for Zero-Shot Counting
Yifei Qian, Zhongliang Guo, Bowen Deng, Chun Tong Lei, Shuai Zhao, Chun Pong Lau , Xiaopeng Hong, Michael P. Pound; Proceedings of the Computer Vision and Patter n Recognition Conference (CVPR), 2025, pp. 25336-25345
Zero-shot object counting aims to count instances of arbitrary object categories specified by text descriptions. Existing methods typically rely on vision-langu age models like CLIP, but often exhibit limited sensitivity to text prompts. We

present T2ICount, a diffusion-based framework that leverages rich prior knowledge and fine-grained visual understanding from pretrained diffusion models. While one-step denoising ensures efficiency, it leads to weakened text sensitivity. To address this challenge, we propose a Hierarchical Semantic Correction Module that progressively refines text-image feature alignment, and a Representational Regional Coherence Loss that provides reliable supervision signals by leveraging the cross-attention maps extracted from the denoising U-Net. Furthermore, we observe that current benchmarks mainly focus on majority objects in images, potentially masking models' text sensitivity. To address this, we contribute a challenging re-annotated subset of FSC147 for better evaluation of text-guided counting ability. Extensive experiments demonstrate that our method achieves superior performance across different benchmarks. Code is available at https://github.com/cha15yq/T2ICount

****

RealEdit: Reddit Edits As a Large-scale Empirical Dataset for Image Transformations

Peter Sushko, Ayana Bharadwaj, Zhi Yang Lim, Vasily Ilin, Ben Caffee, Dongping Chen, Mohammadreza Salehi, Cheng-Yu Hsieh, Ranjay Krishna; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13403-13413

Existing image editing models struggle to meet realworld demands; despite excelling in academic benchmarks, we are yet to see them adopted to solve real user needs. The datasets that power these models use artificial edits, lacking the scale and ecological validity necessary to address the true diversity of user requests. In response, we introduce REALEDIT, a large-scale image editing dataset with authentic user requests and human-made edits sourced from Reddit. REALEDIT contains a test set of 9.3K examples the community can use to evaluate models on real user requests. Our results show that existing models fall short on these tasks, implying a need for realistic training data. So, we introduce 48K training examples, with which we train our REALEDIT model. Our model achieves substantial gains--outperforming competitors by up to 165 Elo points in human judgment and 92% relative improvement on the automated VIEScore metric on our test set. We deploy our model back on Reddit, testing it on new requests, and receive positive feedback. Beyond image editing, we explore REALEDIT 's potential in detecting edited images by partnering with a deepfake detection non-profit. Finetuning their model on REALEDIT data improves its F1-score by 14 percentage points, underscoring the dataset's value for broad, impactful applications.

****

VideoScene: Distilling Video Diffusion Model to Generate 3D Scenes in One Step

Hanyang Wang, Fangfu Liu, Jiawei Chi, Yueqi Duan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16475-16485

Recovering 3D scenes from sparse views is a challenging task due to its inherent ill-posed problem. Conventional methods have developed specialized solutions (e.g., geometry regularization or feed-forward deterministic model) to mitigate the issue. However, they still suffer from performance degradation by minimal overlap across input views with insufficient visual information. Fortunately, recent video generative models show promise in addressing this challenge as they are capable of generating video clips with plausible 3D structures. Powered by large pretrained video diffusion models, some pioneering research start to explore the potential of video generative prior and create 3D scenes from sparse views. Despite impressive improvements, they are limited by slow inference time and the lack of 3D constraint, leading to inefficiencies and reconstruction artifacts that do not align with real-world geometry structure. In this paper, we propose VideoScene to distill the video diffusion model to generate 3D scenes in one step, aiming to build an efficient and effective tool to bridge the gap from video to 3D. Specifically, we design a 3D-aware leap flow distillation strategy to leap over time-consuming redundant information and train a dynamic denoising policy network to adaptively determine the optimal leap timestep during inference. Extensive experiments demonstrate that our VideoScene achieves faster and superior 3D scene generation results than previous video diffusion models, highlighting its potential as an efficient tool for future video to 3D applications.

****
3D-HGS: 3D Half-Gaussian Splatting
Haolin Li, Jinyang Liu, Mario Sznaier, Octavia Camps; Proceedings of the Compute r Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10996-11005
Photo-realistic image rendering from 3D scene reconstruction has advanced signif icantly with neural rendering techniques. Among these, 3D Gaussian Splatting (3D -GS) outperforms Neural Radiance Fields (NeRFs) in quality and speed but struggl es with shape and color discontinuities. We propose 3D Half-Gaussian (3D-HGS) ke rnels as a plug-and-play solution to address these limitations. Our experiments show that 3D-HGS enhances existing 3D-GS methods, achieving state-of-the-art ren dering quality without compromising speed. More demos and code are available at https://lihaolin88.github.io/CVPR-2025-3DHGS
****
FG^2: Fine-Grained Cross-View Localization by Fine-Grained Feature Matching
Zimin Xia, Alexandre Alahi; Proceedings of the Computer Vision and Pattern Recog nition Conference (CVPR), 2025, pp. 6362-6372
We propose a novel fine-grained cross-view localization method that estimates th e 3 Degrees of Freedom pose of a ground-level image in an aerial image of the su rroundings by matching fine-grained features between the two images. The pose is estimated by aligning a point plane generated from the ground image with a poin t plane sampled from the aerial image. To generate the ground points, we first m ap ground image features to a 3D point cloud. Our method then learns to select f eatures along the height dimension to pool the 3D points to a Bird's-Eye-View (B EV) plane. This selection enables us to trace which feature in the ground image contributes to the BEV representation. Next, we sample a set of sparse matches f rom computed point correspondences between the two point planes and compute thei r relative pose using Procrustes alignment. Compared to the previous state-of-th e-art, our method reduces the mean localization error by 28% on the VIGOR cross- area test set. Qualitative results show that our method learns semantically cons istent matches across ground and aerial views through weakly supervised learning from the camera pose.
****
ReNeg: Learning Negative Embedding with Reward Guidance
Xiaomin Li, Yixuan Liu, Takashi Isobe, Xu Jia, Qinpeng Cui, Dong Zhou, Dong Li, You He, Huchuan Lu, Zhongdao Wang, Emad Barsoum; Proceedings of the Computer Vis ion and Pattern Recognition Conference (CVPR), 2025, pp. 23636-23645
In text-to-image (T2I) generation applications, negative embeddings have proven to be a simple yet effective approach for enhancing generation quality. Typicall y, these negative embeddings are derived from user-defined negative prompts, whi ch, while being functional, are not necessarily optimal. In this paper, we intro duce ReNeg, an end-to-end method designed to learn improved Negative embeddings guided by a Reward model. We employ a reward feedback learning framework and int egrate classifier-free guidance (CFG) into the training process, which was previ ously utilized only during inference, thus enabling the effective learning of ne gative embeddings. We also propose two strategies for learning both global and p er-sample negative embeddings. Extensive experiments show that the learned negat ive embedding significantly outperforms null-text and handcrafted counterparts, achieving substantial improvements in human preference alignment. Additionally, the negative embedding learned within the same text embedding space exhibits str ong generalization capabilities. For example, using the same CLIP text encoder, the negative embedding learned on SD1.5 can be seamlessly transferred to text-to -image or even text-to-video models such as ControlNet, ZeroScope, and VideCraft er2, resulting in consistent performance improvements across the board. Code is available at https://github.com/AMD-AIG-AIMA/ReNeg.
****
Scale Efficient Training for Large Datasets
Qing Zhou, Junyu Gao, Qi Wang; Proceedings of the Computer Vision and Pattern Re cognition Conference (CVPR), 2025, pp. 20458-20467
The rapid growth of dataset scales has been a key driver in advancing deep learn ing research. However, as dataset scale increases, the training process becomes

increasingly inefficient due to the presence of low-value samples, including exc
essive redundant samples, overly challenging samples, and inefficient easy sampl
es that contribute little to model improvement. To address this challenge, we pr
opose Scale Efficient Training (SeTa) for large datasets, a dynamic sample pruni
ng approach that losslessly reduces training time. To remove low-value samples,
SeTa first performs random pruning to eliminate redundant samples, then clusters
 the remaining samples according to their learning difficulty measured by loss.
Building upon this clustering, a sliding window strategy is employed to progress
ively remove both overly challenging and inefficient easy clusters following an
easy-to-hard curriculum. We conduct extensive experiments on large-scale synthet
ic datasets, including ToCa, SS1M, and ST+MJ, each containing over 3 million sam
ples. SeTa reduces training costs by up to 50% while maintaining or improving pe
rformance, with minimal degradation even at 70% cost reduction. Furthermore, exp
eriments on various scale real datasets across various backbones (including CNNs
, Transformers, and Mambas) and diverse tasks (instruction tuning, multi-view st
ereo, geo-localization, composed image retrieval, referring image segmentation)
demonstrate the powerful effectiveness and universality of our approach. Code is
 available at https://github.com/mrazhou/SeTa.
****

Distilled Prompt Learning for Incomplete Multimodal Survival Prediction
Yingxue Xu, Fengtao Zhou, Chenyu Zhao, Yihui Wang, Can Yang, Hao Chen; Proceedin
gs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5
102-5111
The integration of multimodal data including pathology images and gene profiles
is widely applied in precise survival prediction. Despite recent advances in mul
timodal survival models, collecting complete modalities for multimodal fusion st
ill poses a significant challenge, hindering their application in clinical setti
ngs. Current approaches tackling incomplete modalities often fall short, as they
 typically compensate for only a limited part of the knowledge of missing modali
ties. To address this issue, we propose a Distilled Prompt Learning framework (D
isPro) to utilize the strong robustness of Large Language Models (LLMs) to missi
ng modalities, which employs two-stage prompting for compensation of comprehensi
ve information for missing modalities. In the first stage, Unimodal Prompting (U
niPro) distills the knowledge distribution of each modality, preparing for suppl
ementing modality-specific knowledge of the missing modality in the subsequent s
tage. In the second stage, Multimodal Prompting (MultiPro) leverages available m
odalities as prompts for LLMs to infer the missing modality, which provides moda
lity-common information. Simultaneously, the unimodal knowledge acquired in the
first stage is injected into multimodal inference to compensate for the modality
-specific knowledge of the missing modality. Extensive experiments covering vari
ous missing scenarios demonstrated the superiority of the proposed method. The c
ode is available at https://github.com/Innse/DisPro.
****

Decoder Gradient Shield: Provable and High-Fidelity Prevention of Gradient-Based
 Box-Free Watermark Removal
Haonan An, Guang Hua, Zhengru Fang, Guowen Xu, Susanto Rahardja, Yuguang Fang; P
roceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 202
5, pp. 13424-13433
The intellectual property of deep image-to-image models can be protected by the
so-called box-free watermarking. It uses an encoder and a decoder, respectively,
 to embed into and extract from the model's output images invisible copyright ma
rks. Prior works have improved watermark robustness, focusing on the design of b
etter watermark encoders. In this paper, we reveal an overlooked vulnerability o
f the unprotected watermark decoder which is jointly trained with the encoder an
d can be exploited to train a watermark removal network. To defend against such
an attack, we propose the decoder gradient shield (DGS) as a protection layer in
 the decoder API to prevent gradient-based watermark removal with a closed-form
solution. The fundamental idea is inspired by the classical adversarial attack,
but is utilized for the first time as a defensive mechanism in the box-free mode
l watermarking. We then demonstrate that DGS can reorient and rescale the gradie

nt directions of watermarked queries and stop the watermark remover's training l
oss from converging to the level without DGS, while retaining decoder output ima
ge quality. Experimental results verify the effectiveness of the proposed method
. Code of paper is available at https://github.com/haonanAN309/CVPR-2025-Officia
l-Implementation-Decoder-Gradient-Shield.
****

MotionPro: A Precise Motion Controller for Image-to-Video Generation
Zhongwei Zhang, Fuchen Long, Zhaofan Qiu, Yingwei Pan, Wu Liu, Ting Yao, Tao Mei
; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR),
2025, pp. 27957-27967
Animating images with interactive motion control has garnered popularity for ima
ge-to-video (I2V) generation. Modern approaches typically rely on large Gaussian
 kernels to extend motion trajectories as condition without explicitly defining
movement region, leading to coarse motion control and failing to disentangle obj
ect and camera moving. To alleviate these, we present MotionPro, a precise motio
n controller that novelly leverages region-wise trajectory and motion mask to re
gulate fine-grained motion synthesis and identify target motion category (i.e.,
object or camera moving), respectively. Technically, MotionPro first estimates t
he flow maps on each training video via a tracking model, and then samples the r
egion-wise trajectories to simulate inference scenario. Instead of extending flo
w through large Gaussian kernels, our region-wise trajectory approach enables mo
re precise control by directly utilizing trajectories within local regions, ther
eby effectively characterizing fine-grained movements. A motion mask is simultan
eously derived from the predicted flow maps to capture the holistic motion dynam
ics of the movement regions. To pursue natural motion control, MotionPro further
 strengthens video denoising by incorporating both region-wise trajectories and
motion mask through feature modulation. More remarkably, we meticulously constru
ct a benchmark, i.e., MC-Bench, with 1.1K user-annotated image-trajectory pairs,
 for the evaluation of both fine-grained and object-level I2V motion control. Ex
tensive experiments conducted on WebVid-10M and MC-Bench demonstrate the effecti
veness of MotionPro. Please refer to our project page for more results: https://
zhw-zhang.github.io/MotionPro-page/.
****

Goku: Flow Based Video Generative Foundation Models
Shoufa Chen, Chongjian Ge, Yuqi Zhang, Yida Zhang, Fengda Zhu, Hao Yang, Hongxia
ng Hao, Hui Wu, Zhichao Lai, Yifei Hu, Ting-Che Lin, Shilong Zhang, Fu Li, Chuan
 Li, Xing Wang, Yanghua Peng, Peize Sun, Ping Luo, Yi Jiang, Zehuan Yuan, Bingyu
e Peng, Xiaobing Liu; Proceedings of the Computer Vision and Pattern Recognition
 Conference (CVPR), 2025, pp. 23516-23527
This paper introduces Goku, a state-of-the-art family of joint image-and-video g
eneration models leveraging rectified flow Transformers to achieve industry-lead
ing performance. We detail the foundational elements enabling high-quality visua
l generation, including the data curation pipeline, model architecture design, f
low formulation, and advanced infrastructure for efficient and robust large-scal
e training. The Goku models demonstrate superior performance in both qualitative
 and quantitative evaluations, setting new benchmarks across major tasks. Specif
ically, Goku achieves 0.76 on GenEval and 83.65 on DPG-Bench for text-to-image g
eneration, and 84.85 on VBench for text-to-video tasks. We believe that this wor
k provides valuable insights and practical advancements for the research communi
ty in developing joint image-and-video generation models.
****

Learning Conditional Space-Time Prompt Distributions for Video Class-Incremental
 Learning
Xiaohan Zou, Wenchao Ma, Shu Zhao; Proceedings of the Computer Vision and Patter
n Recognition Conference (CVPR), 2025, pp. 4862-4873
Recent advancements in prompt-based learning have significantly advanced image a
nd video class-incremental learning. However, the prompts learned by these metho
ds often fail to capture the diverse and informative characteristics of videos,
and struggle to generalize effectively to future tasks and classes. To address t
hese challenges, this paper proposes modeling the distribution of space-time pro

mpts conditioned on the input video using a diffusion model. This generative app
roach allows the proposed model to naturally handle the diverse characteristics
of videos, leading to more robust prompt learning and enhanced generalization ca
pabilities. Additionally, we develop a simple yet effective mechanism to transfe
r the token relationship modeling capabilities of pre-trained image transformers
 to spatio-temporal modeling in videos. Our approach has been thoroughly evaluat
ed across four established benchmarks, showing remarkable improvements over exis
ting state-of-the-art methods in video class-incremental learning.
****

Convex Combination Star Shape Prior for Data-driven Image Semantic Segmentation
Xinyu Zhao, Jun Xie, Shengzhe Chen, Jun Liu; Proceedings of the Computer Vision
and Pattern Recognition Conference (CVPR), 2025, pp. 14068-14077
Multi-center star shape is a prevalent object shape feature, which has proven ef
fective in model-based image segmentation methods. However, the shape field func
tion induced by the multi-center star shape is non-smooth, and directly applying
 it to the data-driven image segmentation network architecture design may lead t
o instability in backpropagation. This paper proposes a convex combination star
(CCS) shape, possessing multi-center star shape properties, and has the advantag
e of effectively controlling the shape of the region through a smooth field func
tion. The sufficient condition of the proposed CCS shape can be combined into th
e image segmentation neural network structure design through the bridge between
the variational segmentation model and the activation function of the data-drive
n method. Taking Segment Anything Model (SAM) and its improved version as backbo
ne networks, we have shown that the segmentation network architecture with CCS s
hape properties can greatly improve the accuracy of segmentation results.
****

Hyperbolic Safety-Aware Vision-Language Models
Tobia Poppi, Tejaswi Kasarla, Pascal Mettes, Lorenzo Baraldi, Rita Cucchiara; Pr
oceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025
, pp. 4222-4232
Addressing the retrieval of unsafe content from vision-language models such as C
LIP is an important step towards real-world integration. Current efforts have re
lied on unlearning techniques that try to erase the model's knowledge of unsafe
concepts. While effective in reducing unwanted outputs, unlearning limits the mo
del's capacity to discern between safe and unsafe content. In this work, we intr
oduce a novel approach that shifts from unlearning to an awareness paradigm by l
everaging the inherent hierarchical properties of the hyperbolic space. We propo
se to encode safe and unsafe content as an entailment hierarchy, where both are
placed in different regions of hyperbolic space. Our HySAC, Hyperbolic Safety-Aw
are CLIP, employs entailment loss functions to model the hierarchical and asymme
trical relations between safe and unsafe image-text pairs. This modelling - inef
fective in standard vision-language models due to their reliance on Euclidean em
beddings - endows the model with awareness of unsafe content, enabling it to ser
ve as both a multimodal unsafe classifier and a flexible content retriever, with
 the option to dynamically redirect unsafe queries toward safer alternatives or
retain the original output. Extensive experiments show that our approach not onl
y enhances safety recognition, but also establishes a more adaptable and interpr
etable framework for content moderation in vision-language models. Our source co
de is available at: https://github.com/aimagelab/HySAC
****

WISH: Weakly Supervised Instance Segmentation using Heterogeneous Labels
Hyeokjun Kweon, Kuk-Jin Yoon; Proceedings of the Computer Vision and Pattern Rec
ognition Conference (CVPR), 2025, pp. 25377-25387
Instance segmentation traditionally relies on dense pixel-level annotations, mak
ing it costly and labor-intensive. To alleviate this burden, weakly supervised i
nstance segmentation utilizes cost-effective weak labels, such as image-level ta
gs, points, and bounding boxes. However, existing approaches typically focus on
a single type of weak label, overlooking the cost-efficiency potential of combin
ing multiple types. In this paper, we introduce WISH, a novel heterogeneous fram
ework for weakly supervised instance segmentation that integrates diverse weak l

abel types within a single model. WISH unifies heterogeneous labels by leveragin g SAM's prompt latent space through a multi-stage matching strategy, effectively compensating for the lack of spatial information in class tags. Extensive exper iments on Pascal VOC and COCO demonstrate that our framework not only surpasses existing homogeneous weak supervision methods but also achieves superior results in heterogeneous settings with equivalent annotation costs.
****

SinGS: Animatable Single-Image Human Gaussian Splats with Kinematic Priors
Yufan Wu, Xuanhong Chen, Wen Li, Shunran Jia, Hualiang Wei, Kairui Feng, Jialian g Chen, Yuhan Li, Ang He, Weimin Zhang, Bingbing Ni, Wenjun Zhang; Proceedings o f the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5571-5580
Despite significant advances in accurately estimating geometry in contemporary s ingle-image 3D human reconstruction, creating a high-quality, efficient, and ani matable 3D avatar remains an open challenge. Two key obstacles persist: incompl ete observation and inconsistent 3D priors. To address these challenges, we prop ose SinGS, aiming to achieve high-quality and efficient animatable 3D avatar rec onstruction. At the heart of SinGS are two key components: Kinematic Human Diffu sion and Geometry-Preserving 3D Gaussain Splatting. The former is a foundational human model that samples within pose space to generate a highly 3D-consistent a nd high-quality sequence of human images, inferring unseen viewpoints and provid ing kinematic priors. The latter is a system that reconstructs a compact, high-quality 3D avatar even under imperfect priors, achieved through a novel semantic Laplacian regularization and a geometry-preserving density control strategy tha t enable precise and compact assembly of 3D primitives. Extensive experiments de monstrate that SinGS enables lifelike, animatable human reconstructions, maintai ning both high quality and inference efficiency (up to 70FPS).
****

Parameter-efficient Fine-tuning in Hyperspherical Space for Open-vocabulary Sema ntic Segmentation
Zelin Peng, Zhengqin Xu, Zhilin Zeng, Yu Huang, Yaoming Wang, Wei Shen; Proceedi ngs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15009-15020
Open-vocabulary semantic segmentation seeks to label each pixel in an image with arbitrary text descriptions. Vision-language foundation models, especially CLIP , have recently emerged as powerful tools for acquiring open-vocabulary capabili ties. However, fine-tuning CLIP to equip it with pixel-level prediction ability often suffers three issues: 1) high computational cost, 2) misalignment between the two inherent modalities of CLIP, and 3) degraded generalization ability on u nseen categories. To address these issues, we propose \alg, a symmetrical parame ter-efficient fine-tuning (PEFT) strategy conducted in hyperspherical space for both of the two CLIP modalities. Specifically, the PEFT strategy is achieved by a series of efficient block-diagonal learnable transformation matrices and a dua l cross-relation communication module among all learnable matrices. Since the PE FT strategy is conducted symmetrically to the two CLIP modalities, the misalignm ent between them is mitigated. Furthermore, we apply an additional constraint to PEFT on the CLIP text encoder according to the hyperspherical energy principle, i.e., minimizing hyperspherical energy during fine-tuning preserves the intrins ic structure of the original parameter space, to prevent the destruction of the generalization ability offered by the CLIP text encoder. Extensive evaluations a cross various benchmarks show that H-CLIP achieves new SOTA open-vocabulary sema ntic segmentation results while only requiring updating approximately 4% of the total parameters of CLIP.
****

Relative Pose Estimation through Affine Corrections of Monocular Depth Priors
Yifan Yu, Shaohui Liu, Rémi Pautrat, Marc Pollefeys, Viktor Larsson; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 167 06-16716
Monocular depth estimation (MDE) models have undergone significant advancements over recent years. Many MDE models aim to predict affine-invariant relative dept

h from monocular images, while recent developments in large-scale training and v
ision foundation models enable reasonable estimation of metric (absolute) depth.
 However, effectively leveraging these predictions for geometric vision tasks, i
n particular relative pose estimation, remains relatively under explored. While
depths provide rich constraints for cross-view image alignment, the intrinsic no
ise and ambiguity from the monocular depth priors present practical challenges t
o improving upon classic keypoint-based solutions. In this paper, we develop thr
ee solvers for relative pose estimation that explicitly account for independent
affine (scale and shift) ambiguities, covering both calibrated and uncalibrated
conditions. We further propose a hybrid estimation pipeline that combines our pr
oposed solvers with classic point-based solvers and epipolar constraints. We fin
d that the affine correction modeling is beneficial to not only the relative dep
th priors but also, surprisingly, the "metric" ones. Results across multiple dat
asets demonstrate large improvements of our approach over classic keypoint-based
 baselines and PnP-based solutions, under both calibrated and uncalibrated setup
s. We also show that our method improves consistently with different feature mat
chers and MDE models, and can further benefit from very recent advances on both
modules.
****

Zero-1-to-A: Zero-Shot One Image to Animatable Head Avatars Using Video Diffusio
n

Zhenglin Zhou, Fan Ma, Hehe Fan, Tat-Seng Chua; Proceedings of the Computer Visi
on and Pattern Recognition Conference (CVPR), 2025, pp. 15941-15952

Animatable head avatar generation typically requires extensive data for training
. To reduce the data requirements, a natural solution is to leverage existing da
ta-free static avatar generation methods, such as pre-trained diffusion models w
ith score distillation sampling (SDS), which align avatars with pseudo ground-tr
uth outputs from the diffusion model. However, directly distilling 4D avatars fr
om video diffusion often leads to over-smooth results due to spatial and tempora
l inconsistencies in the generated video. To address this issue, we propose Zero
-1-to-A, a robust method that synthesizes a spatial and temporal consistency dat
aset for 4D avatar reconstruction using the video diffusion model. Specifically,
 Zero-1-to-A iteratively constructs video datasets and optimizes animatable avat
ars in a progressive manner, ensuring that avatar quality increases smoothly and
 consistently throughout the learning process. This progressive learning involve
s two stages: (1) Spatial Consistency Learning fixes expressions and learns from
 front-to-side views, and (2) Temporal Consistency Learning fixes views and lear
ns from relaxed to exaggerated expressions, generating 4D avatars in a simple-to
-complex manner. Extensive experiments demonstrate that Zero-1-to-A improves fid
elity, animation quality, and rendering speed compared to existing diffusion-bas
ed methods, providing a solution for lifelike avatar creation. Code is publicly
available at: https://github.com/ZhenglinZhou/Zero-1-to-A.
****

Occlusion-aware Text-Image-Point Cloud Pretraining for Open-World 3D Object Reco
gnition

Khanh Nguyen, Ghulam Mubashar Hassan, Ajmal Mian; Proceedings of the Computer Vi
sion and Pattern Recognition Conference (CVPR), 2025, pp. 16965-16975

Recent open-world representation learning approaches have leveraged CLIP to enab
le zero-shot 3D object recognition. However, performance on real point clouds wi
th occlusions still falls short due to unrealistic pretraining settings. Additio
nally, these methods incur high inference costs because they rely on Transformer
's attention modules. In this paper, we make two contributions to address these
limitations. First, we propose occlusion-aware text-image-point cloud pretrainin
g to reduce the training-testing domain gap. From 52K synthetic 3D objects, our
framework generates nearly 630K partial point clouds for pretraining, consistent
ly improving real-world recognition performances of existing popular 3D networks
. Second, to reduce computational requirements, we introduce DuoMamba, a two-str
eam linear state space model tailored for point clouds. By integrating two space
-filling curves with 1D convolutions, DuoMamba effectively models spatial depend
encies between point tokens, offering a powerful alternative to Transformer. Whe

n pretrained with our framework, DuoMamba surpasses current state-of-the-art met
hods while reducing latency and FLOPs, highlighting the potential of our approac
h for real-world applications. Our code and data are available at ndkhanh360.git
hub.io/project_occtip.
****

Conical Visual Concentration for Efficient Large Vision-Language Models
Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang
Cao, Conghui He, Jiaqi Wang, Feng Wu, Dahua Lin; Proceedings of the Computer Vis
ion and Pattern Recognition Conference (CVPR), 2025, pp. 14593-14603
In large vision-language models (LVLMs), images serve as inputs that carry a wea
lth of information. As the idiom "A picture is worth a thousand words" implies,
representing a single image in current LVLMs can require hundreds or even thousa
nds of tokens. This results in significant computational costs, which grow quadr
atically as input image resolution increases, thereby severely impacting the eff
iciency. Previous approaches have attempted to reduce the number of image tokens
 either before or within the early layers of LVLMs. However, these strategies in
evitably result in the loss of crucial image information. To address this challe
nge, we conduct an empirical study revealing that all visual tokens are necessar
y for LVLMs in the shallow layers, and token redundancy progressively increases
in the deeper layers.To this end, we propose ViCo, a conical-style visual concen
tration strategy for LVLMs to boost their efficiency in both training and infere
nce with neglectable performance loss. Specifically, we partition the LVLM into
several stages and drop part of the image tokens at the end of each stage with a
 pre-defined ratio. The dropping is based on a lightweight similarity calculatio
n with a negligible time overhead. Extensive experiments demonstrate that ViCo c
an achieve over 40% training time reduction and 55% inference FLOPs acceleration
 on leading LVLMs like LLaVA-NeXT, maintaining comparable multi-modal performanc
e. Besides, ViCo can also serve as a plug-and-play strategy to accelerate infere
nce in a free way, with better performance and lower inference cost than counter
parts.
****

Good, Cheap, and Fast: Overfitted Image Compression with Wasserstein Distortion
Jona Ballé, Luca Versari, Emilien Dupont, Hyunjik Kim, Matthias Bauer; Proceedin
gs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2
3259-23268
Inspired by the success of generative image models, recent work on learned image
 compression increasingly focuses on better probabilistic models of the natural
image distribution, leading to excellent image quality. This, however, comes at
the expense of a computational complexity that is several orders of magnitude hi
gher than today's commercial codecs, and thus prohibitive for most practical app
lications. With this paper, we demonstrate that by focusing on modeling visual p
erception rather than the data distribution, we can achieve a very good trade-of
f between visual quality and bit rate similar to "generative" compression models
 such as HiFiC, while requiring less than 1% of the multiply-accumulate operatio
ns (MACs) for decompression. We do this by optimizing C3, an overfitted image co
dec, for Wasserstein Distortion (WD), and evaluating the image reconstructions w
ith a human rater study, showing that WD clearly outperforms LPIPS as an optimiz
ation objective. The study also reveals that WD outperforms other perceptual met
rics such as LPIPS, DISTS, and MS-SSIM as a predictor of human ratings, remarkab
ly achieving over 94% Pearson correlation with Elo scores.
****

Period-LLM: Extending the Periodic Capability of Multimodal Large Language Model
Yuting Zhang, Hao Lu, Qingyong Hu, Yin Wang, Kaishen Yuan, Xin Liu, Kaishun Wu;
Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 20
25, pp. 29237-29247
Periodic or quasi-periodic phenomena reveal intrinsic characteristics in various
 natural processes, such as weather patterns, movement behaviors, traffic flows,
 and biological signals. Given that these phenomena span multiple modalities, th
e capabilities of Multimodal Large Language Models (MLLMs) offer promising poten
tial to effectively capture and understand their complex nature. However, curren

t MLLMs struggle with periodic tasks due to limitations in: 1) lack of temporal modelling and 2) conflict between short and long periods. This paper introduces Period-LLM, a multimodal large language model designed to enhance the performance of periodic tasks across various modalities, and constructs a benchmark of various difficulty for evaluating the cross-modal periodic capabilities of large models. Specially, We adopt an "Easy to Hard Generalization" paradigm, starting with relatively simple text-based tasks and progressing to more complex visual and multimodal tasks, ensuring that the model gradually builds robust periodic reasoning capabilities. Additionally, we propose a "Resisting Logical Oblivion" optimization strategy to maintain periodic reasoning abilities during semantic alignment. Extensive experiments demonstrate the superiority of the proposed Period-LLM over existing MLLMs in periodic tasks. The code is available at https://github.com/keke-nice/Period-LLM.

****

V2X-R: Cooperative LiDAR-4D Radar Fusion with Denoising Diffusion for 3D Object Detection

Xun Huang, Jinlong Wang, Qiming Xia, Siheng Chen, Bisheng Yang, Xin Li, Cheng Wang, Chenglu Wen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27390-27400

Current Vehicle-to-Everything (V2X) systems have significantly enhanced 3D object detection using LiDAR and camera data. However, they face performance degradation in adverse weather. Weather-robust 4D radar, with Doppler velocity and additional geometric information, offers a promising solution to this challenge. To this end, we present V2X-R, the first simulated V2X dataset incorporating LiDAR, camera, and 4D radar modalities. V2X-R contains 12,079 scenarios with 37,727 frames of LiDAR and 4D radar point clouds, 150,908 images, and 170,859 annotated 3D vehicle bounding boxes. Subsequently, we propose a novel cooperative LiDAR-4D radar fusion pipeline for 3D object detection and implement it with multiple fusion strategies. To achieve weather-robust detection, we additionally propose a Multi-modal Denoising Diffusion (MDD) module in our fusion pipeline. MDD utilizes weather-robust 4D radar feature as a condition to guide the diffusion model in denoising noisy LiDAR features.Experiments show that our LiDAR-4D radar fusion pipeline demonstrates superior performance in the V2X-R dataset. Over and above this, our MDD module further improved the foggy/snowy performance of the basic fusion model by up to 5.73%/6.70% and barely disrupting normal performance. The dataset and code will be publicly available.

****

Multi-Modal Synergistic Implicit Image Enhancement for Efficient Optical Flow Estimation

Weichen Dai, Hexing Wu, Xiaoyang Weng, Yuxin Zheng, Yuhang Ming, Wanzeng Kong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2173-2182

As a fundamental visual task, optical flow estimation has widespread applications in computer vision. However, it faces significant challenges under adverse lighting conditions, where low texture and noise make accurate optical flow estimation particularly difficult.In this paper, we propose an optical flow method that employs implicit image enhancement through multi-modal synergistic training. To supplement the scene information missing in the original low-quality image, we utilize a high-low frequency feature enhancement network. The enhancement network is implicitly guided by multi-modal data and the specific subsequent tasks, enabling the model to learn multi-modal knowledge that enhances feature information suitable for optical flow estimation during inference. By using RGBD multi-modal data, the proposed method avoids the reliance on the images captured from the same view, a common limitation in traditional image enhancement methods.During training, the encoded features extracted from the enhanced images are synergistically supervised by features from the RGBD fusion as well as by the optical flow task.Experiments conducted on both synthetic and real datasets demonstrate that the proposed method significantly improves performance on public datasets.

****

TAROT: Towards Essentially Domain-Invariant Robustness with Theoretical Justific

ation
Dongyoon Yang, Jihu Lee, Yongdai Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25780-25789
Robust domain adaptation against adversarial attacks is a critical research area that aims to develop models capable of maintaining consistent performance across diverse and challenging domains. In this paper, we derive a new generalization bound for robust risk on the target domain using a novel divergence measure specifically designed for robust domain adaptation. Building upon this, we propose a new algorithm named TAROT, which is designed to enhance both domain adaptability and robustness. Through extensive experiments, TAROT not only surpasses state-of-the-art methods in accuracy and robustness but also significantly enhances domain generalization and scalability by effectively learning domain-invariant features. In particular, TAROT achieves superior performance on the challenging DomainNet dataset, demonstrating its ability to learn domain-invariant representations that generalize well across different domains, including unseen ones. These results highlight the broader applicability of our approach in real-world domain adaptation scenarios.
****

Foundations of the Theory of Performance-Based Ranking
Sébastien Piérard, Anaïs Halin, Anthony Cioppa, Adrien Deliege, Marc Van Droogenbroeck; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14293-14302
Ranking entities such as algorithms, devices, methods, or models based on their performances, while accounting for application-specific preferences, is a challenge. To address this challenge, we establish the foundations of a universal theory for performance-based ranking. First, we introduce a rigorous framework built on top of both the probability and order theories. Our new framework encompasses the elements necessary to (1) manipulate performances as mathematical objects, (2) express which performances are worse than or equivalent to others, (3) model tasks through a variable called satisfaction, (4) consider properties of the evaluation, (5) define scores, and (6) specify application-specific preferences through a variable called importance. On top of this framework, we propose the first axiomatic definition of performance orderings and performance-based rankings. Then, we introduce a universal parametric family of scores, called ranking scores, that can be used to establish rankings satisfying our axioms, while considering application-specific preferences. Finally, we show, in the case of two-class classification, that the family of ranking scores encompasses well-known performance scores, including the accuracy, the true positive rate (recall, sensitivity), the true negative rate (specificity), the positive predictive value (precision), and F1. However, we also show that some other scores commonly used to compare classifiers are unsuitable to derive performance orderings satisfying the axioms.
****

Unveiling the Mist over 3D Vision-Language Understanding: Object-centric Evaluation with Chain-of-Analysis
Jiangyong Huang, Baoxiong Jia, Yan Wang, Ziyu Zhu, Xiongkun Linghu, Qing Li, Song-Chun Zhu, Siyuan Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24570-24581
Existing 3D vision-language (3D-VL) benchmarks fall short in evaluating 3D-VL models, creating a "mist" that obscures rigorous insights into model capabilities and 3D-VL tasks. This mist persists due to three key limitations. First, flawed test data, like ambiguous referential text in the grounding task, can yield incorrect and unreliable test results. Second, oversimplified metrics such as simply averaging accuracy per question answering (QA) pair, cannot reveal true model capability due to their vulnerability to language variations. Third, existing benchmarks isolate the grounding and QA tasks, disregarding the underlying coherence that QA should be based on solid grounding capabilities. To unveil the "mist", we propose Beacon3D, a benchmark for 3D-VL grounding and QA tasks, delivering a perspective shift in the evaluation of 3D-VL understanding. Beacon3D features (i) high-quality test data with precise and natural language, (ii) object-centric

evaluation with multiple tests per object to ensure robustness, and (iii) a novel chain-of-analysis paradigm to address language robustness and model performance coherence across grounding and QA. Our evaluation of state-of-the-art 3D-VL models on Beacon3D reveals that (i) object-centric evaluation elicits true model performance and particularly weak generalization in QA; (ii) grounding-QA coherence remains fragile in current 3D-VL models, and (iii) incorporating large language models (LLMs) to 3D-VL models, though as a prevalent practice, hinders grounding capabilities and has yet to elevate QA capabilities. We hope Beacon3D and our comprehensive analysis could benefit the 3D-VL community towards faithful developments.

****

Generating Multimodal Driving Scenes via Next-Scene Prediction

Yanhao Wu, Haoyang Zhang, Tianwei Lin, Lichao Huang, Shujie Luo, Rui Wu, Congpei Qiu, Wei Ke, Tong Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6844-6853

Generative models in Autonomous Driving (AD) enable diverse scenario creation, yet existing methods fall short by only capturing a limited range of modalities, restricting the capability of generating controllable scenes for comprehensive evaluation of AD systems. In this paper, we introduce a multimodal generation framework that incorporates four major data modalities, including a novel addition of the map modality. With tokenized modalities, our scene sequence generation framework autoregressively predicts each scene while managing computational demands through a two-stage approach. The Temporal AutoRegressive (TAR) component captures inter-frame dynamics for each modality, while the Ordered AutoRegressive (OAR) component aligns modalities within each scene by sequentially predicting tokens in a fixed order. To maintain coherence between map and ego-action modalities, we introduce the Action-aware Map Alignment (AMA) module, which applies a transformation based on the ego-action to maintain coherence between these two modalities. Our framework effectively generates complex, realistic driving scenes over extended sequences, ensuring multimodal consistency and offering fine-grained control over scene elements. Project page: https://yanhaowu.github.io/UMGen/

****

BIGS: Bimanual Category-agnostic Interaction Reconstruction from Monocular Videos via 3D Gaussian Splatting

Jeongwan On, Kyeonghwan Gwak, Gunyoung Kang, Junuk Cha, Soohyun Hwang, Hyein Hwang, Seungryul Baek; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17437-17447

Reconstructing 3Ds of hand-object interaction (HOI) is a fundamental problem that can find numerous applications. Despite recent advances, there is no comprehensive pipeline yet for bimanual class-agnostic interaction reconstruction from a monocular RGB video, where two hands and an unknown object are interacting with each other. Previous works tackled the limited hand-object interaction case, where object templates are pre-known or only one hand is involved in the interaction. The bimanual interaction reconstruction exhibits severe occlusions introduced by complex interactions between two hands and an object. To solve this, we first introduce BIGS (Bimanual Interaction 3D Gaussian Splatting), a method that reconstructs 3D Gaussians of hands and an unknown object from a monocular video. To robustly obtain object Gaussians avoiding severe occlusions, we leverage prior knowledge of pre-trained diffusion model with score distillation sampling (SDS) loss, to reconstruct unseen object parts. For hand Gaussians, we exploit the 3D priors of hand model (i.e., MANO) and share a single Gaussian for two hands to effectively accumulate hand 3D information, given limited views. To further consider the 3D alignment between hands and objects, we include the interacting-subjects optimization step during Gaussian optimization. Our method achieves the state-of-the-art accuracy on two challenging datasets, in terms of 3D hand pose estimation (MPJPE), 3D object reconstruction (CDh, CDo, F10), and rendering quality (PSNR, SSIM, LPIPS), respectively.

****

APT: Adaptive Personalized Training for Diffusion Models with Limited Data

JungWoo Chae, Jiyoon Kim, JaeWoong Choi, Kyungyul Kim, Sangheum Hwang; Proceedin

gs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28619-28628
Personalizing diffusion models using limited data presents significant challenges, including overfitting, loss of prior knowledge, and degradation of text alignment. Overfitting leads to shifts in the noise prediction distribution, disrupting the denoising trajectory and causing the model to lose semantic coherence. In this paper, we propose Adaptive Personalized Training, a novel framework that mitigates overfitting by employing adaptive training strategies and regularizing the model's internal representations during fine-tuning. APT consists of three key components: (1) Adaptive Training Adjustment, which introduces an overfitting indicator to detect the degree of overfitting at each time step bin and applies adaptive data augmentation and adaptive loss weighting based on this indicator; (2)Representation Stabilization, which regularizes the mean and variance of intermediate feature maps to prevent excessive shifts in noise prediction; and (3) Attention Alignment for Prior Knowledge Preservation, which aligns the cross-attention maps of the fine-tuned model with those of the pretrained model to maintain prior knowledge and semantic coherence. Through extensive experiments, we demonstrate that APT effectively mitigates overfitting, preserves prior knowledge, and outperforms existing methods in generating high-quality, diverse images with limited reference data.
****
Symmetry Strikes Back: From Single-Image Symmetry Detection to 3D Generation
Xiang Li, Zixuan Huang, Anh Thai, James M. Rehg; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 743-752
Symmetry is a ubiquitous and fundamental property in the visual world, serving as a critical cue for perception and structure interpretation. This paper investigates the detection of 3D reflection symmetry from a single RGB image, and reveals its significant benefit on single-image 3D generation. We introduce Reflect3D, a scalable, zero-shot symmetry detector capable of robust generalization to diverse and real-world scenarios. Inspired by the success of foundation models, our method scales up symmetry detection with a transformer-based architecture. We also leverage generative priors from multi-view diffusion models to address the inherent ambiguity in single-view symmetry detection. Extensive evaluations on various data sources demonstrate that Reflect3D establishes a new state-of-the-art in single-image symmetry detection. Furthermore, we show the practical benefit of incorporating detected symmetry into single-image 3D generation pipelines through a symmetry-aware optimization process. The integration of symmetry significantly enhances the structural accuracy, cohesiveness, and visual fidelity of the reconstructed 3D geometry and textures, advancing the capabilities of 3D content creation.
****
Frequency-Biased Synergistic Design for Image Compression and Compensation
Jiaming Liu, Qi Zheng, Zihao Liu, Yilian Zhong, Peiye Liu, Tao Liu, Shusong Xu, Yanheng Lu, Sicheng Li, Dimin Niu, Yibo Fan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12820-12829
Compression artifacts removal (CAR), an effective post-processing method to reduce compression distortion in edge-side codecs, demonstrates remarkable results by utilizing convolutional neural networks (CNNs) on high computational power cloud side. Traditional image compression reduces redundancy in the frequency domain, and we observed that CNNs also exhibit a bias in frequency domain when handling compression distortions. However, no prior research leverages this frequency bias to design compression methods tailored to CAR CNNs, or vice versa. In this paper, we present a synergistic design that bridges the gap between image compression and learnable compensation for CAR. Our investigation reveals that different compensation networks have varying effects on low and high-frequencies. Building upon these insights, we propose a pioneering redesign of the quantization process, a fundamental component in lossy image compression, to more effectively compress low-frequency information. Additionally, we devise a novel compensation framework that applies different neural networks for reconstructing different frequencies, incorporating a basis attention block to prioritize intentionally dro

pped low-frequency information, thereby enhancing the overall compensation. We i
nstantiate two compensation networks based on this synergistic design and conduc
t extensive experiments on three image compression standards, demonstrating that
 our approach significantly reduces bitrate consumption while delivering high pe
rceptual quality.
****

PosterMaker: Towards High-Quality Product Poster Generation with Accurate Text R
endering
Yifan Gao, Zihang Lin, Chuanbin Liu, Min Zhou, Tiezheng Ge, Bo Zheng, Hongtao Xi
e; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR),
 2025, pp. 8083-8093
Product posters, which integrate subject, scene, and text, are crucial promotion
al tools for attracting customers. Creating such posters using modern image gene
ration methods is valuable, while the main challenge lies in accurately renderin
g text, especially for complex writing systems like Chinese, which contains over
 10,000 individual characters. In this work, we identify the key to precise text
 rendering as constructing a character-discriminative visual feature as a contro
l signal. Based on this insight, we propose a robust character-wise representati
on as control and we develop TextRenderNet, which achieves a high text rendering
 accuracy of over 90%. Another challenge in poster generation is maintaining the
 fidelity of user-specific products. We address this by introducing SceneGenNet,
 an inpainting-based model, and propose subject fidelity feedback learning to fu
rther enhance fidelity. Based on TextRenderNet and SceneGenNet, we present Poste
rMaker, an end-to-end generation framework. To optimize PosterMaker efficiently,
 we implement a two-stage training strategy that decouples text rendering and ba
ckground generation learning. Experimental results show that PosterMaker outperf
orms existing baselines by a remarkable margin, which demonstrates its effective
ness.
****

Sparse Voxels Rasterization: Real-time High-fidelity Radiance Field Rendering
Cheng Sun, Jaesung Choe, Charles Loop, Wei-Chiu Ma, Yu-Chiang Frank Wang; Procee
dings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp
. 16187-16196
We propose an efficient radiance field rendering algorithm that incorporates a r
asterization process on adaptive sparse voxels without neural networks or 3D Gau
ssians. There are two key contributions coupled with the proposed system. The fi
rst is to adaptively and explicitly allocate sparse voxels to different levels o
f detail within scenes, faithfully reproducing scene details with 65536^3 grid r
esolution while achieving high rendering frame rates. Second, we customize a ras
terizer for efficient adaptive sparse voxels rendering. We render voxels in the
correct depth order by using ray direction-dependent Morton ordering, which avoi
ds the well-known popping artifact found in Gaussian splatting. Our method impro
ves the previous neural-free voxel model by over 4db PSNR and more than 10x FPS
speedup, achieving state-of-the-art comparable novel-view synthesis results. Add
itionally, our voxel representation is seamlessly compatible with grid-based 3D
processing techniques such as Volume Fusion, Voxel Pooling, and Marching Cubes,
enabling a wide range of future extensions and applications. Code at https://git
hub.com/NVlabs/svraster
****

Rethinking Personalized Aesthetics Assessment: Employing Physique Aesthetics Ass
essment as An Exemplification
Haobin Zhong, Shuai He, Anlong Ming, Huadong Ma; Proceedings of the Computer Vis
ion and Pattern Recognition Conference (CVPR), 2025, pp. 2935-2944
The Personalized Aesthetics Assessment (PAA) aims to accurately predict an indiv
idual's unique perception of aesthetics. With the surging demand for customizati
on, PAA enables applications to generate personalized outcomes by aligning with
individual aesthetic preferences. The prevailing PAA paradigm involves two stage
s: pre-training and fine-tuning, but it faces three inherent challenges: 1) The
model is pre-trained using datasets of the Generic Aesthetics Assessment (GAA),
but the collective preferences of GAA lead to conflicts in individualized aesthe

tic predictions. 2) The scope and stage of personalized surveys are related to both the user and the assessed object; however, the prevailing personalized surveys fail to adequately address assessed objects' characteristics. 3) During application usage, the cumulative multimodal feedback from an individual holds great value that should be considered for improving the PAA model but unfortunately attracts insufficient attention. To address the aforementioned challenges, we introduce a new PAA paradigm called PAA+, which is structured into three distinct stages: pre-training, fine-tuning, and continual learning. Furthermore, to better reflect individual differences, we employ a familiar and intuitive application, physique aesthetics assessment (PhysiqueAA), to validate the PAA+ paradigm. We propose a dataset called PhysiqueAA50K, consisting of over 50,000 annotated physique images. Furthermore, we develop a PhysiqueAA framework (PhysiqueFrame) and conduct a large-scale benchmark, achieving state-of-the-art (SOTA) performance. Our research is expected to provide an innovative roadmap and application for the PAA community. The code and dataset are available in here.

****

You See it, You Got it: Learning 3D Creation on Pose-Free Videos at Scale

Baorui Ma, Huachen Gao, Haoge Deng, Zhengxiong Luo, Tiejun Huang, Lulu Tang, Xinlong Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2016-2029

Recent 3D generation models typically rely on limited-scale 3D `gold-labels' or 2D diffusion priors for 3D content creation. However, their performance is upper-bounded by constrained 3D priors due to the lack of scalable learning paradigms. In this work, we present See3D, a visual-conditional multi-view diffusion model trained on large-scale Internet videos for open-world 3D creation. The model aims to Get 3D knowledge by solely Seeing the visual contents from the vast and rapidly growing video data --- You See it, You Got it. To achieve this, we first scale up the training data using a proposed data curation pipeline that automatically filters out multi-view inconsistencies and insufficient observations from source videos. This results in a high-quality, richly diverse, large-scale dataset of multi-view images, termed WebVi3D, containing 320M frames from 16M video clips. Nevertheless, learning generic 3D priors from videos without explicit 3D geometry or camera pose annotations is nontrivial, and annotating poses for web-scale videos is prohibitively expensive. To eliminate the need for pose conditions, we introduce an innovative visual-condition - a purely 2D-inductive visual signal generated by adding time-dependent noise to the masked video data. Finally, we introduce a novel visual-conditional 3D generation framework by integrating See3D into a warping-based pipeline for high-fidelity 3D generation. Our numerical and visual comparisons on single and sparse reconstruction benchmarks show that See3D, trained on cost-effective and scalable video data, achieves notable zero-shot and open-world generation capabilities, markedly outperforming models trained on costly and constrained 3D datasets. Additionally, our model naturally supports other image-conditioned 3D creation tasks, such as 3D editing, without further fine-tuning. Please refer to our project page at: https://vision.baai.ac.cn/see3d.

****

MambaIC: State Space Models for High-Performance Learned Image Compression

Fanhu Zeng, Hao Tang, Yihua Shao, Siyu Chen, Ling Shao, Yan Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18041-18050

A high-performance image compression algorithm is crucial for real-time information transmission across numerous fields. Despite rapid progress in image compression, computational inefficiency and poor redundancy modeling still pose significant bottlenecks, limiting practical applications. Inspired by the effectiveness of state space models (SSMs) in capturing long-range dependencies, we leverage SSMs to address computational inefficiency in existing methods and improve image compression from multiple perspectives. In this paper, we systematically analyze the advantages of SSMs for better integration and propose an enhanced image compression approach through refined context modeling, which we term MambaIC. Specifically, we explore context modeling to adaptively refine the representation of

hidden states. Additionally, we introduce window-based local attention into channel-spatial entropy modeling to reduce potential spatial redundancy during compression, thereby increasing efficiency. Comprehensive qualitative and quantitative results validate the effectiveness and efficiency of our approach, particularly for high-resolution image compression. Code is released at https://github.com/AuroraZengfh/MambaIC.

****

SCAP: Transductive Test-Time Adaptation via Supportive Clique-based Attribute Prompting

Chenyu Zhang, Kunlun Xu, Zichen Liu, Yuxin Peng, Jiahuan Zhou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30032-30041

Vision-language models (VLMs) encounter considerable challenges when adapting to domain shifts stemming from changes in data distribution. Test-time adaptation (TTA) has emerged as a promising approach to enhance VLM performance under such conditions. In practice, test data often arrives in batches, leading to increasing interest in the transductive TTA setting. However, existing TTA methods primarily focus on individual test samples, overlooking crucial cross-sample correlations within a batch. While recent ViT-based TTA methods have introduced batch-level adaptation, they remain suboptimal for VLMs due to inadequate integration of the text modality. To address these limitations, we propose a novel transductive TTA framework, Supportive Clique-based Attribute Prompting (SCAP), which effectively combines visual and textual information to enhance adaptation by generating fine-grained attribute prompts across test batches. SCAP first forms supportive cliques of test samples in an unsupervised manner based on visual similarity and learns an attribute prompt for each clique, capturing shared attributes critical for adaptation. For each test sample, SCAP aggregates attribute prompts from its associated cliques, providing enriched contextual information. To ensure adaptability over time, we incorporate a retention module that dynamically updates attribute prompts and their associated attributes as new data arrives. Comprehensive experiments across multiple benchmarks demonstrate that SCAP outperforms existing state-of-the-art methods, significantly advancing VLM generalization under domain shifts. Our code is available at https://github.com/zhoujiahuan1991/CVPR2025-SCAP.

****

Instant Gaussian Stream: Fast and Generalizable Streaming of Dynamic Scene Reconstruction via Gaussian Splatting

Jinbo Yan, Rui Peng, Zhiyan Wang, Luyang Tang, Jiayu Yang, Jie Liang, Jiahao Wu, Ronggang Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16520-16531

Building Free-Viewpoint Videos in a streaming manner offers the advantage of rapid responsiveness compared to offline training methods, greatly enhancing user experience. However, current streaming approaches face challenges of high per-frame reconstruction time (10s+) and error accumulation, limiting their broader application. In this paper, we propose Instant Gaussian Stream (IGS), a fast and generalizable streaming framework, to address these issues. First, we introduce a generalized Anchor-driven Gaussian Motion Network, which projects multi-view 2D motion features into 3D space, using anchor points to drive the motion of all Gaussians. This generalized Network generates the motion of Gaussians for each target frame in the time required for a single inference. Second, we propose a Key-frame-guided Streaming Strategy that refines each key frame, enabling accurate reconstruction of temporally complex scenes while mitigating error accumulation. We conducted extensive in-domain and cross-domain evaluations, demonstrating that our approach can achieve streaming with a average per-frame reconstruction time of 2s+, alongside a enhancement in view synthesis quality.

****

Locality-Aware Zero-Shot Human-Object Interaction Detection

Sanghyun Kim, Deunsol Jung, Minsu Cho; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20190-20200

Recent methods for zero-shot Human-Object Interaction (HOI) detection typically

leverage the generalization ability of large Vision-Language Model (VLM), i.e., CLIP, on unseen categories, showing impressive results on various zero-shot settings. However, existing methods struggle to adapt CLIP representations for human-object pairs, as CLIP tends to overlook fine-grained information necessary for distinguishing interactions. To address this issue, we devise, LAIN, a novel zero-shot HOI detection framework designed to enhance the locality and interaction awareness of CLIP representations. The locality awareness, which involves capturing fine-grained details and the spatial structure of individual objects, is achieved by aggregating the information and spatial priors of adjacent neighborhood patches. The interaction awareness, which involves identifying whether and how a human is interacting with an object, is achieved by capturing the interaction pattern between the human and the object. By infusing locality and interaction awareness into CLIP representation, LAIN captures detailed information about the human-object pairs. Our extensive experiments on existing benchmarks show that LAIN outperforms previous methods in various zero-shot settings, demonstrating the importance of locality and interaction awareness for effective zero-shot HOI detection.

****

PEACE: Empowering Geologic Map Holistic Understanding with MLLMs

Yangyu Huang, Tianyi Gao, Haoran Xu, Qihao Zhao, Yang Song, Zhipeng Gui, Tengchao Lv, Hao Chen, Lei Cui, Scarlett Li, Furu Wei; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3899-3908

Geologic map, as a fundamental diagram in geology science, provides critical insights into the structure and composition of Earth's subsurface and surface. These maps are indispensable in various fields, including disaster assessment, resource exploration, and civil engineering. Despite their significance, current Multimodal Large Language Models (MLLMs) often fall short in geologic map understanding. This gap is primarily due to the challenging nature of cartographic generalization, which involves handling high-resolution map, managing multiple associated components, and requiring domain-specific knowledge. To quantify this gap, we construct **GeoMap-Bench**, the first-ever benchmark for evaluating MLLMs in geologic map understanding, which assesses the full-scale abilities in extracting, referring, grounding, reasoning, and analyzing. To bridge this gap, we introduce **GeoMap-Agent**, the inaugural agent designed for geologic map understanding, which features three modules: Hierarchical Information Extraction (HIE), Domain Knowledge Injection (DKI), and Prompt-enhanced Question Answering (PEQA). Inspired by the interdisciplinary collaboration among human scientists, an AI expert group acts as consultants, utilizing a diverse tool pool to comprehensively analyze questions. Through comprehensive experiments, GeoMap-Agent achieves an overall score of 0.811 on GeoMap-Bench, significantly outperforming 0.369 of GPT-4o. Our work, em**P**owering g**E**ologic m**A**p holisti**C** und**E**rstanding (**PEACE**) with MLLMs, paves the way for advanced AI applications in geology, enhancing the efficiency and accuracy of geological investigations. The code and data are available at https://github.com/microsoft/PEACE.

****

Tracktention: Leveraging Point Tracking to Attend Videos Faster and Better

Zihang Lai, Andrea Vedaldi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22809-22819

Temporal consistency is critical in video prediction. Traditional methods, such as temporal attention mechanisms and 3D convolutions, often struggle with significant object movements and fail to capture long-range temporal dependencies in dynamic scenes. To address these limitations, we propose the Tracktention Layer, a novel architectural component that explicitly integrates motion information using point tracks -- sequences of corresponding points across frames. By incorporating these motion cues, the Tracktention Layer enhances temporal alignment and effectively handles complex object motions, maintaining consistent feature representations over time. Our approach is computationally efficient and can be seamlessly integrated into existing models, such as Vision Transformers, with minimal modification. Empirical evaluations on standard video estimation benchmarks demonstrate that models augmented with the Tracktention Layer exhibit significantly

improved temporal consistency compared to baseline models.
****
ConceptGuard: Continual Personalized Text-to-Image Generation with Forgetting and Confusion Mitigation
Zirun Guo, Tao Jin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2945-2954
Diffusion customization methods have achieved impressive results with only a minimal number of user-provided images. However, existing approaches customize concepts collectively, whereas real-world applications often require sequential concept integration. This sequential nature can lead to catastrophic forgetting, where previously learned concepts are lost. In this paper, we investigate concept forgetting and concept confusion in the continual customization. To tackle these challenges, we present ConceptGuard, a comprehensive approach that combines shift embedding, concept-binding prompts and memory preservation regularization, supplemented by a priority queue which can adaptively update the importance and occurrence order of different concepts. These strategies can dynamically update, unbind and learn the relationship of the previous concepts, thus alleviating concept forgetting and confusion. Through comprehensive experiments, we show that our approach outperforms all the baseline methods consistently and significantly in both quantitative and qualitative analyses.
****
Two by Two: Learning Multi-Task Pairwise Objects Assembly for Generalizable Robot Manipulation
Yu Qi, Yuanchen Ju, Tianming Wei, Chi Chu, Lawson L.S. Wong, Huazhe Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17383-17393
3D assembly tasks, such as furniture assembly and component fitting, play a crucial role in daily life and represent essential capabilities for future home robots. Existing benchmarks and datasets predominantly focus on assembling geometric fragments or factory parts, which fall short in addressing the complexities of everyday object interactions and assemblies. To bridge this gap, we present 2BY2, a large-scale annotated dataset for daily pairwise objects assembly, covering 18 fine-grained tasks that reflect real-life scenarios, such as plugging into sockets, arranging flowers in vases, and inserting bread into toasters. 2BY2 dataset includes 1,034 instances and 517 pairwise objects with pose and symmetry annotations, requiring approaches that align geometric shapes while accounting for functional and spatial relationships between objects. Leveraging the 2BY2 dataset, we propose a two-step SE(3) pose estimation method with equivariant features for assembly constraints. Compared to previous shape assembly methods, our approach achieves state-of-the-art performance across all 18 tasks in the 2BY2 dataset. Additionally, robot experiments further validate the reliability and generalization ability of our method for complex 3D assembly tasks. More details and demonstrations can be found at https://tea-lab.github.io/TwoByTwo/.
****
SGFormer: Satellite-Ground Fusion for 3D Semantic Scene Completion
Xiyue Guo, Jiarui Hu, Junjie Hu, Hujun Bao, Guofeng Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11929-11938
Recently, camera-based solutions have been extensively explored for scene semantic completion (SSC). Despite their success in visible areas, existing methods struggle to capture complete scene semantics due to frequent visual occlusions. To address this limitation, this paper presents the first satellite-ground cooperative SSC framework, i.e., SGFormer, exploring the potential of satellite-ground image pairs in the SSC task. Specifically, we propose a dual-branch architecture that encodes orthogonal satellite and ground views in parallel, unifying them into a common domain. Additionally, we design a ground-view guidance strategy that pre-corrects satellite image biases during feature encoding, addressing misalignment between satellite and ground views. Moreover, we develop an adaptive weighting strategy that balances contributions from satellite and ground views. Experiments demonstrate that SGFormer outperforms the state of the art on SemanticKITTI and SSCBench-KITTI-360 datasets. Our code is available on https://github.com

/gxytcrc/SGFormer.
****
MARVEL-40M+: Multi-Level Visual Elaboration for High-Fidelity Text-to-3D Content Creation
Sankalp Sinha, Mohammad Sadil Khan, Muhammad Usama, Shino Sam, Didier Stricker, Sk Aziz Ali, Muhammad Zeshan Afzal; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8105-8116
Generating high-fidelity 3D content from text prompts remains a significant challenge in computer vision due to the limited size, diversity, and annotation depth of the existing datasets. To address this, we introduce MARVEL-40M+, an extensive dataset with 40 million text annotations for over 8.9 million 3D assets aggregated from seven major 3D datasets. Our contribution is a novel multi-stage annotation pipeline that integrates open-source pretrained multi-view VLMs and LLMs to automatically produce multi-level descriptions, ranging from detailed (150-200 words) to concise semantic tags (10-20 words). This structure supports both fine-grained 3D reconstruction and rapid prototyping. Furthermore, we incorporate human metadata from source datasets into our annotation pipeline to add domain-specific information in our annotation and reduce VLM hallucinations. Additionally, we develop MARVEL-FX3D, a two-stage text-to-3D pipeline. We fine-tune Stable Diffusion with our annotations and use a pretrained image-to-3D network to generate 3D textured meshes within 15s. Extensive evaluations show that MARVEL-40M+ significantly outperforms existing datasets in annotation quality and linguistic diversity, achieving win rates of 72.41% by GPT-4 and 73.40% by human evaluators. Project page is available at https://sankalpsinha-cmos.github.io/MARVEL/.
****
Random Conditioning for Diffusion Model Compression with Distillation
Dohyun Kim, Sehwan Park, Geonhee Han, Seung Wook Kim, Paul Hongsuck Seo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18607-18618
Diffusion models generate high-quality images through progressive denoising but are computationally intensive due to large model sizes and repeated sampling. Knowledge distillation--transferring knowledge from a complex teacher to a simpler student model--has been widely studied in recognition tasks, particularly for transferring concepts unseen during student training. However, its application to diffusion models remains underexplored, especially in enabling student models to generate concepts not covered by the training images. In this work, we propose Random Conditioning, a novel approach that pairs noised images with randomly selected text conditions to enable efficient, image-free knowledge distillation. By leveraging this technique, we show that the student can generate concepts unseen in the training images. When applied to conditional diffusion model distillation, our method allows the student to explore the condition space without generating condition-specific images, resulting in notable improvements in both generation quality and efficiency. This promotes resource-efficient deployment of generative diffusion models, broadening their accessibility for both research and real-world applications. Code, models, and datasets are available at: https://dohyun-as.github.io/Random-Conditioning
****
Hierarchical Gaussian Mixture Model Splatting for Efficient and Part Controllable 3D Generation
Qitong Yang, Mingtao Feng, Zijie Wu, Weisheng Dong, Fangfang Wu, Yaonan Wang, Ajmal Mian; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11104-11114
3D content creation has achieved significant progress in terms of both quality and speed. Although current Gaussian Splatting-based methods can produce 3D objects within seconds, they are still limited by complex preprocessing or low controllability. In this paper, we introduce a novel framework designed to efficiently and controllably generate high-resolution 3D models from text promptsor image. Our key insights are three-fold: 1) Hierarchical Gaussian Mixture Model Splatting: We propose an hybrid hierarchical representation to extract fixed number of fine-grained Gaussians with multiscale details from textured object, also establi

sh part-level representation of Gaussians primitives. 2) Mamba with adaptive tre e topology: We present a diffusion mamba with tree-topology to adaptively genera te Gaussians with disordered spatial structures, without the need for complex pr eprocessing and maintain linear complexity generation. 3) Controllable Generatio n: Building on the HGMM tree, we introduce a cascaded diffusion framework combin ing controllable implicit latent generation, which progressively generates condi tion-driven latents, and explicit splatting generation, which transforms latents into high-quality Gaussian primitives. Extensive experiments demonstrate the hi gh fidelity and efficiency of our approach.
****

ERUPT: Efficient Rendering with Unposed Patch Transformer
Maxim V. Shugaev, Vincent Chen, Maxim Karrenbach, Kyle Ashley, Bridget Kennedy, Naresh P. Cuntoor; Proceedings of the Computer Vision and Pattern Recognition Co nference (CVPR), 2025, pp. 6057-6067
This work addresses the problem of novel view synthesis in diverse scenes from s mall collections of RGB images. We propose ERUPT (Efficient Rendering with Unpos ed Patch Transformer) a state-of-the-art scene reconstruction model capable of e fficient scene rendering using unposed imagery. We introduce patch-based queryin g, in contrast to existing pixel-based queries, to reduce the compute required t o render a target view. This makes our model highly efficient both during traini ng and at inference, capable of rendering at 600 fps on commercial hardware. Not ably, our model is designed to use a learned latent camera pose which allows for training using unposed targets in datasets with sparse or inaccurate ground tru th camera pose. We show that our approach can generalize on large real-world dat a and introduce a new benchmark dataset (MSVS-1M) for latent view synthesis usin g street-view imagery collected from Mapillary. In contrast to NeRF and Gaussian Splatting which require dense imagery and precise metadata, ERUPT can render no vel views of arbitrary scenes with as few as five unposed input images. ERUPT ac hieves better rendered image quality than current state-of-the-art methods for u nposed image synthesis tasks, reduces labeled data requirements by 95% and redu ces computational requirements by an order of magnitude, providing efficient nov el view synthesis for diverse real-world scenes.
****

Rethinking End-to-End 2D to 3D Scene Segmentation in Gaussian Splatting
Runsong Zhu, Shi Qiu, Zhengzhe Liu, Ka-Hei Hui, Qianyi Wu, Pheng-Ann Heng, Chi-W ing Fu; Proceedings of the Computer Vision and Pattern Recognition Conference (C VPR), 2025, pp. 3656-3665
Lifting multi-view 2D instance segmentation to a radiance field has proven effec tive to enhance 3D understanding. Existing works rely on direct matching for end -to-end lifting, yielding inferior results, or employ a two-stage solution const rained by complex pre- or post-processing. In this work, we design Unified-Lift, a new end-to-end object-aware lifting approach that aims for high-quality 3D se gmentation based on our object-aware 3D Gaussian representation. To start, we au gment each Gaussian point with a Gaussian-level feature learned using a contrast ive loss to encode instance information. Importantly, we introduce a learnable o bject-level codebook to account for individual objects in the scene for an expli cit object-level understanding and associate the encoded object-level features w ith the Gaussian-level point features for segmentation predictions. While promis ing, achieving effective codebook learning is nontrivial and a naive solution le ads to degraded performance. Hence, we formulate the association learning module and the noisy label filtering module for effective and robust codebook learning . We conduct experiments on three benchmarks LERF-Masked, Replica, and Messy Roo ms. Both qualitative and quantitative results manifest that our Unified-Lift cle arly outperforms existing methods in terms of segmentation quality and time effi ciency.
****

Quad-Pixel Image Defocus Deblurring: A New Benchmark and Model
Hang Chen, Yin Xie, Xiaoxiu Peng, Lihu Sun, Wenkai Su, Xiaodong Yang, Chengming Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5709-5719

Defocus deblurring is a challenging task due to the spatially varying blur. Recent works have shown impressive results in data-driven approaches using dual-pixel (DP) sensors. Quad-pixel (QP) sensors represent an advanced evolution of DP sensors, providing four distinct sub-aperture views in contrast to only two views offered by DP sensors. However, research on QP-based defocus deblurring is scarce. In this paper, we propose a novel end-to-end learning-based approach for defocus deblurring that leverages QP data. To achieve this, we design a QP defocus and all-in-focus image pair acquisition method and provide a QP Defocus Deblurring (QPDD) dataset containing 4,935 image pairs. We then introduce a Local-gate assisted Mamba Network (LMNet), which includes a two-branch encoder and a Simple Fusion Module (SFM) to fully utilize features of sub-aperture views. In particular, our LMNet incorporates a Local-gate assisted Mamba Block (LAMB) that mitigates local pixel forgetting and channel redundancy within Mamba, and effectively captures global and local dependencies. By extending the defocus deblurring task from a DP-based to a QP-based approach, we demonstrate significant improvements in restoring sharp images. Comprehensive experimental evaluations further indicate that our approach outperforms state-of-the-art methods.

****

DocVLM: Make Your VLM an Efficient Reader

Mor Shpigel Nacson, Aviad Aberdam, Roy Ganz, Elad Ben Avraham, Alona Golts, Yair Kittenplon, Shai Mazor, Ron Litman; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29005-29015

Vision-Language Models (VLMs) excel in diverse visual tasks but face challenges in document understanding, which requires fine-grained text processing. While typical visual tasks perform well with low-resolution inputs, reading-intensive applications demand high-resolution, resulting in significant computational overhead. Using OCR-extracted text in VLM prompts partially addresses this issue but underperforms compared to full-resolution counterpart, as it lacks the complete visual context needed for optimal performance.We introduce DocVLM, a method that integrates an OCR-based modality into VLMs to enhance document processing while preserving original weights. Our approach employs an OCR encoder to capture textual content and layout, compressing these into a compact set of learned queries incorporated into the VLM. Comprehensive evaluations across leading VLMs show that DocVLM significantly reduces reliance on high-resolution images for document understanding.In limited-token regimes (448x448), DocVLM with 64 learned queries improves DocVQA results from 56.0% to 86.6% when integrated with InternVL2 and from 84.4% to 91.2% with Qwen2-VL. In LLaVA-OneVision, DocVLM achieves improved results while using 80% less image tokens. The reduced token usage allows processing multiple pages effectively, showing impressive zero-shot results on DUDE and state-of-the-art performance on MP-DocVQA, highlighting DocVLM's potential for applications requiring high-performance and efficiency.

****

Revisiting Source-Free Domain Adaptation: Insights into Representativeness, Generalization, and Variety

Ronghang Zhu, Mengxuan Hu, Weiming Zhuang, Lingjuan Lyu, Xiang Yu, Sheng Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25688-25697

Domain adaptation addresses the challenge where the distribution of target inference data differs from that of the source training data. Recently, data privacy has become a significant constraint, limiting access to the source domain. To mitigate this issue, Source-Free Domain Adaptation (SFDA) methods bypass source domain data by generating source-like data or pseudo-labeling the unlabeled target domain. However, these approaches often lack theoretical grounding. In this work, we provide a theoretical analysis of the SFDA problem, focusing on the general empirical risk of the unlabeled target domain. Our analysis offers a comprehensive understanding of how representativeness, generalization, and variety contribute to controlling the upper bound of target domain empirical risk in SFDA settings. We further explore how to balance this trade-off from three perspectives: sample selection, semantic domain alignment, and a progressive learning framework. These insights inform the design of novel algorithms. Experimental results de

monstrate that our proposed method achieves state-of-the-art performance on thre e benchmark datasets--Office-Home, DomainNet, and VisDA-C--yielding relative imp rovements of 3.2%, 9.1%, and 7.5%, respectively, over the representative SFDA me thod, SHOT.
****

Adaptive Unimodal Regulation for Balanced Multimodal Information Acquisition
Chengxiang Huang, Yake Wei, Zequn Yang, Di Hu; Proceedings of the Computer Visio n and Pattern Recognition Conference (CVPR), 2025, pp. 25854-25863
Sensory training during the early ages is vital for human development. Inspired by this cognitive phenomenon, we observe that the early training stage is also i mportant for the multimodal learning process, where dataset information is rapid ly acquired. We refer to this stage as the prime learning window. However, based on our observation, this prime learning window in multimodal learning is often dominated by information-sufficient modalities, which in turn suppresses the inf ormation acquisition of information-insufficient modalities. To address this iss ue, we propose Information Acquisition Regulation (InfoReg), a method designed t o balance information acquisition among modalities. Specifically, InfoReg slows down the information acquisition process of information-sufficient modalities du ring the prime learning window, which could promote information acquisition of i nformation-insufficient modalities. This regulation enables a more balanced lear ning process and improves the overall performance of the multimodal network. Exp eriments show that InfoReg outperforms related multimodal imbalanced methods acr oss various datasets, achieving superior model performance. The code is availabl e at https://github.com/GeWu-Lab/InfoReg_CVPR2025.
****

Heterogeneous Skeleton-Based Action Representation Learning
Hongsong Wang, Xiaoyan Ma, Jidong Kuang, Jie Gui; Proceedings of the Computer Vi sion and Pattern Recognition Conference (CVPR), 2025, pp. 19154-19164
Skeleton-based human action recognition has received widespread attention in rec ent years due to its diverse range of application scenarios. Due to the differen t sources of human skeletons, skeleton data naturally exhibit heterogeneity. The previous works, however, overlook the heterogeneity of human skeletons and sole ly construct models tailored for homogeneous skeletons. This work addresses the challenge of heterogeneous skeleton-based action representation learning, specif ically focusing on processing skeleton data that varies in joint dimensions and topological structures. The proposed framework comprises two primary components: heterogeneous skeleton processing and unified representation learning. The form er first converts two-dimensional skeleton data into three-dimensional skeleton via an auxiliary network, and then constructs a prompted unified skeleton using skeleton-specific prompts. We also design an additional modality named semantic motion encoding to harness the semantic information within skeletons. The latter module learns a unified action representation using a shared backbone network t hat processes different heterogeneous skeletons. Extensive experiments on the NT U-60, NTU-120, and PKU-MMD II datasets demonstrate the effectiveness of our meth od in various tasks of action understanding. Our approach can be applied to acti on recognition in robots with different humanoid structures.
****

FLARE: Feed-forward Geometry, Appearance and Camera Estimation from Uncalibrated Sparse Views
Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowe i Zhou, Yujun Shen, Gordon Wetzstein; Proceedings of the Computer Vision and Pat tern Recognition Conference (CVPR), 2025, pp. 21936-21947
We present FLARE, a feed-forward model designed to infer high-quality camera pos es and 3D geometry from uncalibrated sparse-view images (i.e., as few as 2-8 inp uts), which is a challenging yet practical setting in real-world applications. O ur solution features a cascaded learning paradigm with camera pose serving as th e critical bridge, recognizing its essential role in mapping 3D structures onto 2D image planes. Concretely, FLARE starts with camera pose estimation, whose res ults condition the subsequent learning of geometric structure and appearance, op timized through the objectives of geometry reconstruction and novel-view synthes

is. Utilizing large-scale public datasets for training, our method delivers state-of-the-art performance in the tasks of pose estimation, geometry reconstruction, and novel view synthesis, while maintaining the inference efficiency (i.e., less than 0.5 seconds).
****

Improving Gaussian Splatting with Localized Points Management
Haosen Yang, Chenhao Zhang, Wenqing Wang, Marco Volino, Adrian Hilton, Li Zhang, Xiatian Zhu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21696-21705
Point management is critical for optimizing 3D Gaussian Splatting models, as point initiation (e.g., via structure from motion) is often distributionally inappropriate. Typically, Adaptive Density Control (ADC) algorithm is adopted, leveraging view-averaged gradient magnitude thresholding for point densification, opacity thresholding for pruning, and regular all-points opacity reset. We reveal that this strategy is limited in tackling intricate/special image regions (e.g., transparent) due to inability of identifying all 3D zones requiring point densification, and lacking an appropriate mechanism to handle ill-conditioned points with negative impacts (e.g., occlusion due to false high opacity).To address these limitations, we propose a Localized Point Management(LPM) strategy, capable of identifying those error-contributing zones in greatest need for both point addition and geometry calibration. Zone identification is achieved by leveraging the underlying multiview geometry constraints, subject to image rendering errors. We apply point densification in the identified zones and then reset the opacity of the points in front of these regions, creating a new opportunity to correct poorly conditioned points. Serving as a versatile plugin, LPM can be seamlessly integrated into existing static 3D and dynamic 4D Gaussian Splatting models with minimal additional cost.Experimental evaluations validate the efficacy of our LPM in boosting a variety of existing 3D/4D models both quantitatively and qualitatively. Notably, LPM improves both static 3DGS and dynamic SpaceTimeGS to achieve state-of-the-art rendering quality while retaining real-time speeds, excelling on challenging datasets such as Tanks & Temples and the Neural 3D Video dataset.
****

GEAL: Generalizable 3D Affordance Learning with Cross-Modal Consistency
Dongyue Lu, Lingdong Kong, Tianxin Huang, Gim Hee Lee; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1680-1690
Identifying affordance regions on 3D objects from semantic cues is essential for robotics and human-machine interaction. However, existing 3D affordance learning methods struggle with generalization and robustness due to limited annotated data and a reliance on 3D backbones focused on geometric encoding, which often lack resilience to real-world noise and data corruption. We propose GEAL, a novel framework designed to enhance the generalization and robustness of 3D affordance learning by leveraging large-scale pre-trained 2D models. We employ a dual-branch architecture with Gaussian splatting to establish consistent mappings between 3D point clouds and 2D representations, enabling realistic 2D renderings from sparse point clouds. A granularity-adaptive fusion module and a 2D-3D consistency alignment module further strengthen cross-modal alignment and knowledge transfer, allowing the 3D branch to benefit from the rich semantics and generalization capacity of 2D models. To holistically assess the robustness, we introduce two new corruption-based benchmarks: PIAD-C and LASO-C. Extensive experiments on public datasets and our benchmarks show that GEAL consistently outperforms existing methods across seen and novel object categories, as well as corrupted data, demonstrating robust and adaptable affordance predictions. The code and datasets are publicly available.
****

Dynamic Derivation and Elimination: Audio Visual Segmentation with Enhanced Audio Semantics
Chen Liu, Liying Yang, Peike Li, Dadong Wang, Lincheng Li, Xin Yu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3131-3141
Sound-guided object segmentation has drawn considerable attention for its potent

ial to enhance multimodal perception. Previous methods primarily focus on develo
ping advanced architectures to facilitate effective audio-visual interactions, w
ithout fully addressing the inherent challenges posed by audio natures, i.e., (
1) feature confusion due to the overlapping nature of audio signals, and (2) aud
io-visual matching difficulty from the varied sounds produced by the same object
. To address these challenges, we propose Dynamic Derivation and Elimination (DD
ESeg): a novel audio-visual segmentation framework. Specifically, to mitigate fe
ature confusion, DDESeg reconstructs the semantic content of the mixed audio sig
nal by enriching the distinct semantic information of each individual source, de
riving representations that preserve the unique characteristics of each sound. T
o reduce the matching difficulty, we introduce a discriminative feature learning
module, which enhances the semantic distinctiveness of generated audio represen
tations. Considering that not all derived audio representations directly corresp
ond to visual features (e.g., off-screen sounds), we propose a dynamic eliminati
on module to filter out non-matching elements. This module facilitates targeted
interaction between sounding regions and relevant audio semantics. By scoring th
e interacted features, we identify and filter out irrelevant audio information,
ensuring accurate audio-visual alignment. Comprehensive experiments demonstrate
that our framework achieves superior performance in AVS datasets. Our code will
be publicly available.
****

AnyMap: Learning a General Camera Model for Structure-from-Motion with Unknown D
istortion in Dynamic Scenes

Andrea Porfiri Dal Cin, Georgi Dikov, Jihong Ju, Mohsen Ghafoorian; Proceedings
of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1667
4-16684

Current learning-based Structure-from-Motion (SfM) methods struggle with videos
of dynamic scenes from wide-angle cameras. We present AnyMap, a differentiable S
fM framework that jointly addresses image distortion and motion estimation. By l
earning a general implicit camera model without predefined parameters, AnyMap ef
fectively handles lens distortion, estimating multi-view consistent 3D geometry,
camera poses, and (un)projection functions. To resolve the ambiguity where moti
on estimation can compensate for undistortion errors and vice versa, we introduc
e a low-dimensional motion representation consisting of a set of learnable basis
trajectories, interpolated to produce regularized motion estimates. Experimenta
l results show that our method produces accurate camera poses, excels in camera
calibration and image rectification, and enables high-quality novel view synthes
is. Our low-dimensional motion representation effectively disentangles undistort
ion with motion estimation, outperforming existing methods.
****

ESC: Erasing Space Concept for Knowledge Deletion

Tae-Young Lee, Sundong Park, Minwoo Jeon, Hyoseok Hwang, Gyeong-Moon Park; Proce
edings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, p
p. 5010-5019

As concerns regarding privacy in deep learning continue to grow, individuals are
increasingly apprehensive about the potential exploitation of their personal kn
owledge in trained models. Despite several research efforts to address this, the
y often fail to consider the real-world demand from users for complete knowledge
erasure. Furthermore, our investigation reveals that existing methods have a ri
sk of leaking personal knowledge through embedding features. To address these is
sues, we introduce a novel concept of Knowledge Deletion (KD), an advanced task
that considers both concerns, and provides an appropriate metric, named Knowledg
e Retention score (KR), for assessing knowledge retention in feature space. To a
chieve this, we propose a novel training-free erasing approach named Erasing Spa
ce Concept (ESC), which restricts the important subspace for the forgetting know
ledge by eliminating the relevant activations in the feature. In addition, we su
ggest ESC with Training (ESC-T), which uses a learnable mask to better balance t
he trade-off between forgetting and preserving knowledge in KD. Our extensive ex
periments on various datasets and models demonstrate that our proposed methods a
chieve the fastest and state-of-the-art performance. Notably, our methods are ap

plicable to diverse forgetting scenarios, such as facial domain setting, demonst rating the generalizability of our methods. The code is available at https://git hub.com/KU-VGI/ESC.
****

Language Guided Concept Bottleneck Models for Interpretable Continual Learning
Lu Yu, Haoyu Han, Zhe Tao, Hantao Yao, Changsheng Xu; Proceedings of the Compute r Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14976-14986
Continual learning (CL) aims to enable learning systems to acquire new knowledge constantly without forgetting previously learned information. CL faces the chal lenge of mitigating catastrophic forgetting while maintaining interpretability a cross tasks.Most existing CL methods focus primarily on preserving learned knowl edge to improve model performance. However, as new information is introduced, th e interpretability of the learning process becomes crucial for understanding the evolving decision-making process, yet it is rarely explored. In this paper, we introduce a novel framework that integrates language-guided Concept Bottleneck M odels (CBMs) to address both challenges. Our approach leverages the Concept Bott leneck Layer, aligning semantic consistency with CLIP models to learn human-unde rstandable concepts that can generalize across tasks. By focusing on interpretab le concepts, our method not only enhances the model's ability to retain knowledg e over time but also provides transparent decision-making insights. We demonstra te the effectiveness of our approach by achieving superior performance on severa l datasets, outperforming state-of-the-art methods with an improvement of up to 3.06% in final average accuracy on ImageNet-subset. Additionally, we offer conce pt visualizations for model predictions, further advancing the understanding of interpretable continual learning. Code is available at https://github.com/Fisher Cats/CLG-CBM.
****

One-Way Ticket: Time-Independent Unified Encoder for Distilling Text-to-Image Di ffusion Models
Senmao Li, Lei Wang, Kai Wang, Tao Liu, Jiehang Xie, Joost van de Weijer, Fahad Shahbaz Khan, Shiqi Yang, Yaxing Wang, Jian Yang; Proceedings of the Computer Vi sion and Pattern Recognition Conference (CVPR), 2025, pp. 23563-23574
Text-to-Image (T2I) diffusion models have made remarkable advancements in genera tive modeling; however, they face a trade-off between inference speed and image quality, posing challenges for efficient deployment. Existing distilled T2I mode ls can generate high-fidelity images with fewer sampling steps, but often strugg le with diversity and quality, especially in one-step models. From our analysis, we observe redundant computations in the UNet encoders. Our findings suggest th at, for T2I diffusion models, decoders are more adept at capturing richer and mo re explicit semantic information, while encoders can be effectively shared acros s decoders from diverse time steps.Based on these observations, we introduce the first Time-independent Unified Encoder (TiUE) for the student model UNet archit ecture, which is a loop-free image generation approach for distilling T2I diffus ion models. Using a one-pass scheme, TiUE shares encoder features across multipl e decoder time steps, enabling parallel sampling and significantly reducing infe rence time complexity. In addition, we incorporate a KL divergence term to regul arize noise prediction, which enhances the perceptual realism and diversity of t he generated images. Experimental results demonstrate that TiUE outperforms stat e-of-the-art methods, including LCM, SD-Turbo, and SwiftBrushv2, producing more diverse and realistic results while maintaining the computational efficiency.
****

Domain Adaptive Diabetic Retinopathy Grading with Model Absence and Flowing Data
Wenxin Su, Song Tang, Xiaofeng Liu, Xiaojing Yi, Mao Ye, Chunxiao Zu, Jiahao Li, Xiatian Zhu; Proceedings of the Computer Vision and Pattern Recognition Confere nce (CVPR), 2025, pp. 28337-28346
Domain shift (the difference between source and target domains) poses a signific ant challenge in clinical applications, e.g., Diabetic Retinopathy (DR) grading. Despite considering certain clinical requirements, like source data privacy, co nventional transfer methods are predominantly model-centered and often struggle to prevent model-targeted attacks. In this paper, we address a challenging Onlin

e Model-aGnostic Domain Adaptation (OMG-DA) setting, driven by the demands of cl inical environments. This setting is characterized by the absence of the model a nd the flow of target data. To tackle the new challenge, we propose a novel appr oach, Generative Unadversarial ExampleS (GUES), which enables adaptation from a data-centric perspective. Specifically, we first theoretically reformulate conve ntional perturbation optimization in a generative way-- learning a perturbation generation function with a latent input variable. During model instantiation, we leverage a Variational AutoEncoder to express this function. The encoder with t he reparameterization trick predicts the latent input, whilst the decoder is res ponsible for the generation. Furthermore, the saliency map is selected as pseudo perturbation labels. Because it not only captures potential lesions but also th eoretically provides an upper bound on the function input, enabling the identifi cation of the latent variable. Extensive experiments on DR benchmarks with both frozen pre-trained models and trainable models demonstrate the superiority of GU ES, showing robustness even with small batch size. The source code and data are available at https://github.com/tntek/GUES.

****

Temporal Separation with Entropy Regularization for Knowledge Distillation in Sp iking Neural Networks

Kairong Yu, Chengting Yu, Tianqing Zhang, Xiaochen Zhao, Shu Yang, Hongwei Wang, Qiang Zhang, Qi Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8806-8816

Spiking Neural Networks (SNNs), inspired by the human brain, offer significant c omputational efficiency through discrete spike-based information transfer. Despi te their potential to reduce inference energy consumption, a performance gap per sists between SNNs and Artificial Neural Networks (ANNs), primarily due to curre nt training methods and inherent model limitations. While recent research has ai med to enhance SNN learning by employing knowledge distillation (KD) from ANN te acher networks, traditional distillation techniques often overlook the distincti ve spatiotemporal properties of SNNs, thus failing to fully leverage their advan tages. To overcome these challenge, we propose a novel logit distillation method characterized by temporal separation and entropy regularization. This approach improves existing SNN distillation techniques by performing distillation learnin g on logits across different time steps, rather than merely on aggregated output features. Furthermore, the integration of entropy regularization stabilizes mod el optimization and further boosts the performance. Extensive experimental resul ts indicate that our method surpasses prior SNN distillation strategies, whether based on logit distillation, feature distillation, or a combination of both. Ou r project is available at https://github.com/yukairong/TSER.

****

LoRASculpt: Sculpting LoRA for Harmonizing General and Specialized Knowledge in Multimodal Large Language Models

Jian Liang, Wenke Huang, Guancheng Wan, Qu Yang, Mang Ye; Proceedings of the Com puter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26170-26180

While Multimodal Large Language Models (MLLMs) excel at generalizing across moda lities and tasks, effectively adapting them to specific downstream tasks while s imultaneously retaining both general and specialized knowledge remains challengi ng. Although Low-Rank Adaptation (LoRA) is widely used to efficiently acquire sp ecialized knowledge in MLLMs, it introduces substantial harmful redundancy durin g visual instruction tuning, which exacerbates the forgetting of general knowled ge and degrades downstream task performance.To address this issue, we propose Lo RASculpt to eliminate harmful redundant parameters, thereby harmonizing general and specialized knowledge.Specifically, under theoretical guarantees, we introdu ce sparse updates into LoRA to discard redundant parameters effectively. Further more, we propose a Conflict Mitigation Regularizer to refine the update trajecto ry of LoRA, mitigating knowledge conflicts with the pretrained weights.Extensive experimental results demonstrate that even at very high degree of sparsity (\le 5%), our method simultaneously enhances generalization and downstream task perf ormance. This confirms that our approach effectively mitigates the catastrophic forgetting issue and further promotes knowledge harmonization in MLLMs.

****
SEAL: Semantic Attention Learning for Long Video Representation
Lan Wang, Yujia Chen, Du Tran, Vishnu Naresh Boddeti, Wen-Sheng Chu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26192-26201
Long video understanding presents challenges due to the inherent high computational complexity and redundant temporal information. An effective representation for long videos must efficiently process such redundancy while preserving essential contents for downstream tasks. This paper introduces **SE**mantic **A**ttention **L**earning (SEAL), a novel unified representation for long videos. To reduce computational complexity, long videos are decomposed into three distinct types of semantic entities: scenes, objects, and actions, allowing models to operate on a compact set of entities rather than a large number of frames or pixels. To further address redundancy, we propose an attention learning module that balances token relevance with diversity, formulated as a subset selection optimization problem. Our representation is versatile, enabling applications across various long video understanding tasks. Our representation is versatile and applicable across various long video understanding tasks. Extensive experiments demonstrate that SEAL significantly outperforms state-of-the-art methods in video question answering and temporal grounding tasks across diverse benchmarks, including LVBench, MovieChat-1K, and Ego4D.
****
Re-HOLD: Video Hand Object Interaction Reenactment via adaptive Layout-instructed Diffusion Model
Yingying Fan, Quanwei Yang, Kaisiyuan Wang, Hang Zhou, Yingying Li, Haocheng Feng, Errui Ding, Yu Wu, Jingdong Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17550-17560
Current digital human studies focusing on lip-syncing and body movement are no longer sufficient to meet the growing industrial demand, while human video generation techniques that support interacting with real-world environments (e.g., objects) have not been well investigated. Despite human hand synthesis already being an intricate problem, generating objects in contact with hands and their interactions presents an even more challenging task, especially when the objects exhibit obvious variations in size and shape. To tackle these issues, we present a novel video Reenactment framework focusing on Human-Object Interaction (HOI) via an adaptive Layout-instructed Diffusion model (Re-HOLD). Our key insight is to employ specialized layout representation for hands and objects, respectively. Such representations enable effective disentanglement of hand modeling and object adaptation to diverse motion sequences. To further improve the generation quality of HOI, we design an interactive textural enhancement module for both hands and objects by introducing two independent memory banks. We also propose a layout adjustment strategy for the cross-object reenactment scenario to adaptively adjust unreasonable layouts caused by diverse object sizes during inference. Comprehensive qualitative and quantitative evaluations demonstrate that our proposed framework significantly outperforms existing methods. Project page: https://fyycs.github.io/Re-HOLD.
****
Theoretical Insights in Model Inversion Robustness and Conditional Entropy Maximization for Collaborative Inference Systems
Song Xia, Yi Yu, Wenhan Yang, Meiwen Ding, Zhuo Chen, Ling-Yu Duan, Alex C. Kot, Xudong Jiang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8753-8763
By locally encoding raw data into intermediate features, collaborative inference enables end users to leverage powerful deep learning models without exposure of sensitive raw data to cloud servers. However, recent studies have revealed that these intermediate features may not sufficiently preserve privacy, as information can be leaked and raw data can be reconstructed via model inversion attacks (MIAs). Obfuscation-based methods, such as noise corruption, adversarial representation learning, and information filters, enhance the inversion robustness by obfuscating the task-irrelevant redundancy empirically. However, methods for quant

ifying such redundancy remain elusive, and the explicit mathematical relation between this redundancy and worst-case robustness against inversion has not yet been established. To address that, this work first theoretically proves that the conditional entropy of inputs given intermediate features provides a guaranteed lower bound on the reconstruction mean square error (MSE) under any MIA. Then, we derive a differentiable and solvable measure for bounding this conditional entropy based on the Gaussian mixture estimation and propose a conditional entropy maximization (CEM) algorithm to enhance the inversion robustness. Experimental results on four datasets demonstrate the effectiveness and adaptability of our proposed CEM; without compromising the feature utility and computing efficiency, integrating the proposed CEM into obfuscation-based defense mechanisms consistently boosts their inversion robustness, achieving average gains ranging from 12.9% to 48.2%. Code is available at \href https://github.com/xiasong0501/CEM https://github.com/xiasong0501/CEM .

****

Odd-One-Out: Anomaly Detection by Comparing with Neighbors

Ankan Bhunia, Changjian Li, Hakan Bilen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20395-20404

This paper introduces a novel anomaly detection (AD) problem aimed at identifying `odd-looking' objects within a scene by comparing them to other objects present. Unlike traditional AD benchmarks with fixed anomaly criteria, our task detects anomalies specific to each scene by inferring a reference group of regular objects. To address occlusions, we use multiple views of each scene as input, construct 3D object-centric models for each instance from 2D views, enhancing these models with geometrically consistent part-aware representations. Anomalous objects are then detected through cross-instance comparison. We also introduce two new benchmarks, ToysAD-8K and PartsAD-15K as testbeds for future research in this task. We provide a comprehensive analysis of our method quantitatively and qualitatively on these benchmarks.

****

SCFlow2: Plug-and-Play Object Pose Refiner with Shape-Constraint Scene Flow

Qingyuan Wang, Rui Song, Jiaojiao Li, Kerui Cheng, David Ferstl, Yinlin Hu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22045-22054

We introduce SCFlow2, a plug-and-play refinement framework for 6D object pose estimation. Most recent 6D object pose methods rely on refinement to get accurate results. However, most existing refinements either suffer from noises in establishing correspondences, or rely on retraining for novel objects. SCFlow2 is based on the SCFlow model designed for iterative RGB refinement with shape constraint, but formulates the additional depth as a regularization in the iteration via 3D scene flow for RGBD frames. The key design of SCFlow2 is an introduction of geometry constraints into the training of recurrent match network, by combining the rigid-motion embeddings in 3D scene flow and 3D shape prior of the target. We train the refinement network on a combination of dataset Objaverse, GSO and ShapeNet, and demonstrate on BOP datasets with novel objects that, after using our method, the result of most state-of-the-art methods improves significantly, without any retraining or fine-tuning.

****

D^3CTTA: Domain-Dependent Decorrelation for Continual Test-Time Adaption of 3D LiDAR Segmentation

Jichun Zhao, Haiyong Jiang, Haoxuan Song, Jun Xiao, Dong Gong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11864-11874

Adapting pre-trained LiDAR segmentation models to dynamic domain shifts during testing is of paramount importance for the safety of autonomous driving. Most existing methods neglect the influence of domain changes and point density in continual test-time adaption (CTTA), relying on backpropagation and large batch sizes for stability. We approach this problem with three insights: 1) Point clouds at different distances usually have different densities resulting in distribution disparities; 2) The feature distribution of different domains varies, and domai

n-aware parameters can alleviate domain gaps; 3) Features are highly correlated and make segmentation of different labels confusing. To this end, this work presents D^3CTTA, an online backpropagation-free framework for 3D continual test-time adaption for LiDAR segmentation. D^3CTTA consists of a distance-aware prototype learning module to integrate LiDAR-based geometry prior and a domain-dependent decorrelation module to reduce feature correlations among different domains and different categories. Extensive experiments on three benchmarks showcase that our method achieves a state-of-the-art performance compared to both backpropagation-based methods and backpropagation-free methods.
****

Flowing from Words to Pixels: A Noise-Free Framework for Cross-Modality Evolution

Qihao Liu, Xi Yin, Alan Yuille, Andrew Brown, Mannat Singh; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2755-2765

Diffusion models, and their generalization, flow matching, have had a remarkable impact on the field of media generation. Here, the conventional approach is to learn the complex mapping from a simple source distribution of Gaussian noise to the target media distribution. For cross-modal tasks such as text-to-image generation, this same mapping from noise to image is learnt whilst including a conditioning mechanism in the model. One key and thus far relatively unexplored feature of flow matching is that, unlike Diffusion models, they are not constrained for the source distribution to be noise. Hence, in this paper, we propose a paradigm shift, and ask the question of whether we can instead train flow matching models to learn a direct mapping from the distribution of one modality to the distribution of another, thus obviating the need for both the noise distribution and conditioning mechanism. We present a general and simple framework, CrossFlow, for cross-modal flow matching. We show the importance of applying Variational Encoders to the input data, and introduce a method to enable Classifier-free guidance. Surprisingly, for text-to-image, CrossFlow with a vanilla transformer without cross attention slightly outperforms standard flow matching, and we show that it scales better with training steps and model size, while also allowing for interesting latent arithmetic which results in semantically meaningful edits in the output space. To demonstrate the generalizability of our approach, we also show that CrossFlow is on par with or outperforms the state-of-the-art for various cross-modal / intra-modal mapping tasks, viz. image captioning, depth estimation, and image super-resolution. We hope this paper contributes to accelerating progress in cross-modal media generation.
****

FlipSketch: Flipping Static Drawings to Text-Guided Sketch Animations

Hmrishav Bandyopadhyay, Yi-Zhe Song; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28394-28404

Sketch animations offer a powerful medium for visual storytelling, from simple flip-book doodles to professional studio productions. While traditional animation requires teams of skilled artists to draw key frames and in-between frames, existing automation attempts still demand significant artistic effort through precise motion paths or keyframe specification. We present FlipSketch, a system that brings back the magic of flip-book animation -- just draw your idea and describe how you want it to move! Our approach harnesses motion priors from text-to-video diffusion models, adapting them to generate sketch animations through three key innovations: (i) fine-tuning for sketch-style frame generation, (ii) a reference frame mechanism that preserves visual integrity of input sketch through noise refinement, and (iii) a dual-attention composition that enables fluid motion without losing visual consistency. Unlike constrained vector animations, our raster frames support dynamic sketch transformations, capturing the expressive freedom of traditional animation. The result is an intuitive system that makes sketch animation as simple as doodling and describing, while maintaining the artistic essence of hand-drawn animation.
****

Interpretable Generative Models through Post-hoc Concept Bottlenecks

Akshay Kulkarni, Ge Yan, Chung-En Sun, Tuomas Oikarinen, Tsui-Wei Weng; Proceedi

ngs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8162-8171
Concept bottleneck models (CBM) aim to produce inherently interpretable models that rely on human-understandable concepts for their predictions. However, existing approaches to design interpretable generative models based on CBMs are not yet efficient and scalable, as they require expensive generative model training from scratch as well as real images with labor-intensive concept supervision. To address these challenges, we present two novel and low-cost methods to build interpretable generative models through post-hoc techniques and we name our approaches: concept-bottleneck autoencoder (CB-AE) and concept controller (CC). Our proposed approaches enable efficient and scalable training without the need of real data and require only minimal to no concept supervision. Additionally, our methods generalize across modern generative model families including generative adversarial networks and diffusion models. We demonstrate the superior interpretability and steerability of our methods on numerous standard datasets like CelebA, CelebA-HQ, and CUB with large improvements (average 25%) over the prior work, while being 4-15x faster to train. Finally, a large-scale user study is performed to validate the interpretability and steerability of our methods.
****

SketchAgent: Language-Driven Sequential Sketch Generation
Yael Vinker, Tamar Rott Shaham, Kristine Zheng, Alex Zhao, Judith E Fan, Antonio Torralba; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23355-23368
Sketching serves as a versatile tool for externalizing ideas, enabling rapid exploration and visual communication that spans various disciplines. While artificial systems have driven substantial advances in content creation and human-computer interaction, capturing the dynamic and abstract nature of human sketching remains challenging. In this work, we introduce SketchAgent, a language-driven, sequential sketch generation method that enables users to create, modify, and refine sketches through dynamic, conversational interactions. Our approach requires no training or fine-tuning. Instead, we leverage the sequential nature and rich prior knowledge of off-the-shelf multimodal large language models (LLMs). We present an intuitive sketching language, introduced to the model through in-context examples, enabling it to "draw" using string-based actions. These are processed into vector graphics and then rendered to create a sketch on a pixel canvas, which can be accessed again for further tasks. By drawing stroke by stroke, our agent captures the evolving, dynamic qualities intrinsic to sketching. We demonstrate that SketchAgent can generate sketches from diverse prompts, engage in dialogue-driven drawing, and collaborate meaningfully with human users.
****

DRAWER: Digital Reconstruction and Articulation With Environment Realism
Hongchi Xia, Entong Su, Marius Memmel, Arhan Jain, Raymond Yu, Numfor Mbiziwo-Tiapo, Ali Farhadi, Abhishek Gupta, Shenlong Wang, Wei-Chiu Ma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21771-21782
Creating virtual digital replicas from real-world data unlocks significant potential across domains like gaming and robotics. In this paper, we present DRAWER, a novel framework that converts a video of a static indoor scene into a photorealistic and interactive digital environment. Our approach centers on two main contributions: (i) a reconstruction module based on a dual scene representation that reconstructs the scene with fine-grained geometric details, and (ii) an articulation module that identifies articulation types and hinge positions, reconstructs simulatable shapes and appearances and integrates them into the scene. The resulting virtual environment is photorealistic, interactive, and runs in real time, with compatibility for game engines and robotic simulation platforms. We demonstrate the potential of DRAWER by using it to automatically create an interactive game in Unreal Engine and to enable real-to-sim-to-real transfer for robotics applications. Project page: https://xiahongchi.github.io/DRAWER/.
****

GoLF-NRT: Integrating Global Context and Local Geometry for Few-Shot View Synthe

sis
You Wang, Li Fang, Hao Zhu, Fei Hu, Long Ye, Zhan Ma; Proceedings of the Compute r Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21349-21359
Neural Radiance Fields (NeRF) have transformed novel view synthesis by modeling scene-specific volumetric representations directly from images. While generaliza ble NeRF models can generate novel views across unknown scenes by learning laten t ray representations, their performance heavily depends on a large number of mu lti-view observations. However, with limited input views, these methods experien ce significant degradation in rendering quality. To address this limitation, we propose GoLF-NRT: a Global and Local feature Fusion-based Neural Rendering Trans former. GoLF-NRT enhances generalizable neural rendering from few input views by leveraging a 3D transformer with efficient sparse attention to capture global s cene context. In parallel, it integrates local geometric features extracted alon g the epipolar line, enabling high-quality scene reconstruction from as few as 1 to 3 input views. Furthermore, we introduce an adaptive sampling strategy based on attention weights and kernel regression, improving the accuracy of transform er-based neural rendering. Extensive experiments on public datasets show that Go LF-NRT achieves state-of-the-art performance across varying numbers of input vie ws, highlighting the effectiveness and superiority of our approach. Code is avai lable at https://github.com/KLMAV-CUC/GoLF-NRT.
****

Deep Change Monitoring: A Hyperbolic Representative Learning Framework and a Dat aset for Long-term Fine-grained Tree Change Detection
Yante Li, Hanwen Qi, Haoyu Chen, Xinlian Liang, Guoying Zhao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27346-2735 6
In environmental protection, tree monitoring plays an essential role in maintain ing and improving ecosystem health. However, precise monitoring is challenging b ecause existing datasets fail to capture continuous fine-grained changes in tree s due to low-resolution images and high acquisition costs. In this paper, we int roduce UAVTC, a large-scale, long-term, high-resolution dataset collected using UAVs equipped with cameras, specifically designed to detect individual Tree Chan ges (TCs). UAVTC includes rich annotations and statistics based on biological kn owledge, offering a fine-grained view for tree monitoring. To address environmen tal influences and effectively model the hierarchical diversity of physiological TCs, we propose a novel Hyperbolic Siamese Network (HSN) for TC detection, enab ling compact and hierarchical representations of dynamic tree changes. Extensive experiments show that HSN can effectively capture complex hierarchical changes and provide a robust solution for fine-grained TC detection. In addition, HSN ge neralizes well to cross-domain face anti-spoofing task, highlighting its broader significance in AI. We believe our work, combining ecological insights and int erdisciplinary expertise, will benefit the community by offering a new benchmark and innovative AI technologies. Source code is available on https://github.c om/liyantett/Tree-Changes-Detection-with-Siamese-Hyperbolic-network.
****

A Closer Look at Time Steps is Worthy of Triple Speed-Up for Diffusion Model Tra ining
Kai Wang, Mingjia Shi, Yukun Zhou, Zekai Li, Zhihang Yuan, Yuzhang Shang, Xiaoji ang Peng, Hanwang Zhang, Yang You; Proceedings of the Computer Vision and Patter n Recognition Conference (CVPR), 2025, pp. 12934-12944
Training diffusion models is always a computation-intensive task. In this paper, we introduce a novel speed-up method for diffusion model training, called, whic h is based on a closer look at time steps. Our key findings are: i) Time steps c an be empirically divided into acceleration, deceleration, and convergence areas based on the process increment. ii) These time steps are imbalanced, with many concentrated in the convergence area. iii) The concentrated steps provide limite d benefits for diffusion training. To address this, we design an asymmetric samp ling strategy that reduces the frequency of steps from the convergence area whil e increasing the sampling probability for steps from other areas. Additionally, we propose a weighting strategy to emphasize the importance of time steps with r

apid-change process increments. As a plug-and-play and architecture-agnostic app roach, SpeeD consistently achieves 3-times acceleration across various diffusion architectures, datasets, and tasks. Notably, due to its simple design, our appr oach significantly reduces the cost of diffusion model training with minimal ove rhead. Our research enables more researchers to train diffusion models at a lowe r cost.
****

Empowering LLMs to Understand and Generate Complex Vector Graphics
Ximing Xing, Juncheng Hu, Guotao Liang, Jing Zhang, Dong Xu, Qian Yu; Proceeding s of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19 487-19497
The unprecedented advancements in Large Language Models (LLMs) have profoundly i mpacted natural language processing but have yet to fully embrace the realm of s calable vector graphics (SVG) generation. While LLMs encode partial knowledge of SVG data from web pages during training, recent findings suggest that semantica lly ambiguous and tokenized representations within LLMs may result in hallucinat ions in vector primitive predictions. Additionally, LLM training typically lacks modeling and understanding of the rendering sequence of vector paths, which can lead to occlusion between output vector primitives. In this paper, we present L LM4SVG, an initial yet substantial step toward bridging this gap by enabling LLM s to better understand and generate vector graphics. LLM4SVG facilitates a deepe r understanding of SVG components through learnable semantic tokens, which preci sely encode these tokens and their corresponding properties to generate semantic ally aligned SVG outputs. Using a series of learnable semantic tokens, a structu red dataset for instruction following is developed to support comprehension and generation across two primary tasks. Our method introduces a modular architectur e to existing large language models, integrating semantic tags, vector instructi on encoders, fine-tuned commands, and powerful LLMs to tightly combine geometric , appearance, and language information. To overcome the scarcity of SVG-text ins truction data, we developed an automated data generation pipeline that collected our SVGX-SFT Dataset, consisting of high-quality human-designed SVGs and 580k S VG instruction following data specifically crafted for LLM training, which facil itated the adoption of the supervised fine-tuning strategy popular in LLM develo pment. By exploring various training strategies, we developed LLM4SVG, which sig nificantly moves beyond optimized rendering-based approaches and language-model-based baselines to achieve remarkable results in human evaluation tasks. Code, m odel, and data will be released at: https://ximinng.github.io/LLM4SVGProject/
****

PanoGS: Gaussian-based Panoptic Segmentation for 3D Open Vocabulary Scene Unders tanding
Hongjia Zhai, Hai Li, Zhenzhe Li, Xiaokun Pan, Yijia He, Guofeng Zhang; Proceedi ngs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14114-14124
Recently, 3D Gaussian Splatting (3DGS) has shown encouraging performance for ope n vocabulary scene understanding tasks. However, previous methods can not distin guish 3D instance-level information, which usually predicts a heatmap between th e scene feature and text query. In this paper, we propose PanoGS, a novel and ef ficient 3D panoptic open vocabulary scene understanding approach. Technical-wise , to learn accurate 3D language features that can scale to large indoor scenario s, we adopt the pyramid tri-planes to model the latent continuous parametric fea ture space and use a 3D feature decoder to regress the multi-view fused 2D featu re cloud. Besides, we propose language-guided graph cuts that synergistically le verage reconstructed geometry and learned language cues to group 3D Gaussian pri mitives into a set of super-primitives. To obtain 3D consistent instance, we per form graph clustering based segmentation with SAM-guided edge affinity computati on between different super-primitives. Extensive experiments on widely used data sets show better or more competitive performance on 3D panoptic open vocabulary scene understanding.
****

Watermarking One for All: A Robust Watermarking Scheme Against Partial Image The

ft
Gaozhi Liu, Silu Cao, Zhenxing Qian, Xinpeng Zhang, Sheng Li, Wanli Peng; Procee dings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp . 8225-8234
The proliferation of digital images on the Internet has provided unprecedented c onvenience, but also poses significant risks of malicious theft and misuse. Digi tal watermarking has long been researched as an effective tool for copyright pro tection. However, it often falls short when addressing partial image theft, a co mmon yet little-researched issue in practical applications. Most existing scheme s typically require the entire image as input to extract watermarks. However, in practice, malicious users often steal only a portion of the image to create new content. The stolen portion can have arbitrary shape or content, being fused wi th a new background and may have undergone geometric transformations, making it challenging for current methods to extract correctly. To address the issues abov e, we propose WOFA (Watermarking One for All), a robust watermarking scheme agai nst partial image theft. First of all, we define the entire process of partial i mage theft and construct a dataset accordingly. To gain robustness against parti al image theft, we then design a comprehensive distortion layer that incorporate s the process of partial image theft and several common distortions in channel. For easier network convergence, we employ a multi-level network structure on the basis of the commonly used embedder-distortion layer-extractor architecture and adopt a progressive training strategy. Abundant experiments demonstrate that ou r superior performance in the scenario of partial image theft, offering a more r eliable solution for protecting digital images against unauthorized use in pract ical use.
****
ITA-MDT: Image-Timestep-Adaptive Masked Diffusion Transformer Framework for Imag e-Based Virtual Try-On
Ji Woo Hong, Tri Ton, Trung X. Pham, Gwanhyeong Koo, Sunjae Yoon, Chang D. Yoo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 20 25, pp. 28284-28294
This paper introduces ITA-MDT, the Image-Timestep-Adaptive Masked Diffusion Tran sformer Framework for Image-Based Virtual Try-On (IVTON), designed to overcome t he limitations of previous approaches by leveraging the Masked Diffusion Transfo rmer (MDT) for improved handling of both global garment context and fine-grained details. The IVTON task involves seamlessly superimposing a garment from one im age onto a person in another, creating a realistic depiction of the person weari ng the specified garment. Unlike conventional diffusion-based virtual try-on mod els that depend on large pre-trained U-Net architectures, ITA-MDT leverages a li ghtweight, scalable transformer-based denoising diffusion model with a mask late nt modeling scheme, achieving competitive results while reducing computational o verhead. A key component of ITA-MDT is the Image-Timestep Adaptive Feature Aggre gator (ITAFA), a dynamic feature aggregator that combines all of the features fr om the image encoder into a unified feature of the same size, guided by diffusio n timestep and garment image complexity. This enables adaptive weighting of feat ures, allowing the model to emphasize either global information or fine-grained details based on the requirements of the denoising stage. Additionally, the Sali ent Region Extractor (SRE) module is presented to identify complex region of the garment to provide high-resolution local information to the denoising model as an additional condition alongside the global information of the full garment ima ge. This targeted conditioning strategy enhances detail preservation of fine det ails in highly salient garment regions, optimizing computational resources by av oiding unnecessarily processing entire garment image. Comparative evaluations co nfirms that ITA-MDT improves efficiency while maintaining strong performance, re aching state-of-the-art results in several metrics.
****
MultiVENT 2.0: A Massive Multilingual Benchmark for Event-Centric Video Retrieva l
Reno Kriz, Kate Sanders, David Etter, Kenton Murray, Cameron Carpenter, Hannah R ecknor, Jimena Guallar-Blasco, Alexander Martin, Eugene Yang, Benjamin Van Durme

; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24149-24158
Efficiently retrieving and synthesizing information from large-scale multimodal collections has become a critical challenge. However, existing video retrieval datasets suffer from scope limitations, primarily focusing on matching descriptive but vague queries with small collections of professionally edited, English-centric videos. To address this gap, we introduce MultiVENT 2.0, a large-scale, multilingual event-centric video retrieval benchmark featuring a collection of more than 218,000 news videos and over 3,900 queries targeting specific world events. These queries specifically target information found in the visual content, audio, embedded text, and text metadata of the videos, requiring systems leverage all these sources to succeed at the task. Preliminary results show that state-of-the-art vision-language models struggle significantly with this task, and while alternative approaches show promise, they are still insufficient to adequately address this problem. These findings underscore the need for more robust multimodal retrieval systems, as effective video retrieval is a crucial step towards multimodal content understanding and generation.
****
VolFormer: Explore More Comprehensive Cube Interaction for Hyperspectral Image Restoration and Beyond
Dabing Yu, Zheng Gao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28091-28101
Capitalizing on the talent of self-attention in capturing non-local features, Transformer architectures have exhibited remarkable performance in single hyperspectral image restoration. For hyperspectral images, each pixel is located in the hyperspectral image cubes with a large spectral dimension and two spatial dimensions. Although uni-dimensional self-attention, like channel self-attention or spatial self-attention, builds long-range dependencies in spectral or spatial dimensions, they lack more comprehensive interactions across dimensions. To tackle the above drawback, we propose a VolFormer, a volumetric self-attention embedded Transformer network for single hyperspectral image restoration. Specifically, we propose volumetric self-attention (VolSA), which extends the interaction from 2D flat to 3D cube. VolSA can simultaneously model token interaction in the 3D cube, mining the potential correlations between the hyperspectral image cube. An attention decomposition form is proposed to reduce the computational burden of modeling volumetric information. In practical terms, VolSA adapts double similarity matrixes in spatial and channel dimensions to implicitly model 3D context information while transforming the complexity from cubic to quadratic. Additionally, we introduce the explicit spectral location prior to enhance the proposed self-attention. This property allows the target token to perceive global spectral information while simultaneously assigning different levels of attention to tokens at varying wavelength bands. Extensive experiments demonstrate that VolFormer achieves record-high performance on hyperspectral image super-resolution, denoise and classification benchmarks. Particularly, VolSA is portable and achieves inspiring results in hyperspectral classification. The source code is available at https://github.com/yudadabing/VolFormer.
****
Minding Fuzzy Regions: A Data-driven Alternating Learning Paradigm for Stable Lesion Segmentation
Lexin Fang, Yunyang Xu, Xiang Ma, Xuemei Li, Caiming Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10425-10434
Deep learning has achieved significant advancements in medical image segmentation, but existing models still face challenges in accurately segmenting lesion regions. The main reason is that some lesion regions in medical images have unclear boundaries, irregular shapes, and small tissue density differences, leading to label ambiguity. However, the existing model treats all data equally without taking quality differences into account in the training process, resulting in noisy labels negatively impacting model training and unstable feature representations. In this paper, a **d**ata-driven **a**lternating **le**arning (**DALE**) paradigm is proposed to optimize the model's training process, achieving stable and

high-precision segmentation. The paradigm focuses on two key points: (1) reducing the impact of noisy labels, and (2) calibrating unstable representations. To mitigate the negative impact of noisy labels, a loss consistency-based collaborative optimization method is proposed, and its effectiveness is theoretically demonstrated. Specifically, the label confidence parameters are introduced to dynamically adjust the influence of labels of different confidence levels during model training, thus reducing the influence of noise labels. To calibrate the learning bias of unstable representations, a distribution alignment method is proposed. This method restores the underlying distribution of unstable representations, thereby enhancing the discriminative capability of fuzzy region representations. Extensive experiments on various benchmarks and model backbones demonstrate the superiority of the DALE paradigm, achieving an average performance improvement of up to 7.16%.

****

BizGen: Advancing Article-level Visual Text Rendering for Infographics Generation

Yuyang Peng, Shishi Xiao, Keming Wu, Qisheng Liao, Bohan Chen, Kevin Lin, Danqing Huang, Ji Li, Yuhui Yuan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23615-23624

Recently, state-of-the-art text-to-image generation models, such as Flux and Ideogram 2.0, have made significant progress in sentence-level visual text rendering. In this paper, we focus on the more challenging scenarios of article-level visual text rendering and address a novel task of generating high-quality business content, including infographics and slides, based on user provided article-level descriptive prompts and ultra-dense layouts. The fundamental challenges are twofold: significantly longer context lengths and the scarcity of high-quality business content data. In contrast to most previous works that focus on a limited number of sub-regions and sentence-level prompts, ensuring precise adherence to ultra-dense layouts with tens or even hundreds of sub-regions in business content is far more challenging. We make two key technical contributions: (i) the construction of scalable, high-quality business content dataset, i.e.,Infographics-650K, equipped with ultra-dense layouts and prompts by implementing a layer-wise retrieval-augmented infographic generation scheme; and (ii) a layout-guided cross attention scheme, which injects tens of region-wise prompts into a set of cropped region latent space according to the ultra-dense layouts, and refine each sub-regions flexibly during inference using a layout conditional CFG. We demonstrate the strong results of our system compared to previous SOTA systems such as Flux and SD3 on our BizEval prompt set. Additionally, we conduct thorough ablation experiments to verify the effectiveness of each component. We hope our constructed Infographics-650K and BizEval can encourage the broader community to advance the progress of business content generation.

****

MLVU: Benchmarking Multi-task Long Video Understanding

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, Zheng Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13691-13701

The evaluation of Long Video Understanding (LVU) performance poses an important but challenging research problem. Despite previous efforts, the existing video understanding benchmarks are severely constrained by several issues, especially the insufficient lengths of videos, a lack of diversity in video types and evaluation tasks, and the inappropriateness for evaluating LVU performances. To address the above problems, we propose a new benchmark called MLVU (Multi-task Long Video Understanding Benchmark) for the comprehensive and in-depth evaluation of LVU. MLVU presents the following critical values: 1) The substantial and flexible extension of video lengths, which enables the benchmark to evaluate LVU performance across a wide range of durations. 2) The inclusion of various video genres, such as movies, surveillance, egocentric videos, and cartoons, reflects the models' LVU performances in different scenarios. 3) The development of diversified evaluation tasks, which enables a comprehensive examination of MLLMs' key abiliti

es in long-video understanding. The empirical study with 23 latest MLLMs reveals significant room for improvement in today's technique, as all existing methods struggle with most of the evaluation tasks and exhibit severe performance degradation when handling longer videos. Additionally, it suggests that factors such as context length, image-understanding ability, and the choice of LLM backbone can play critical roles in future advancements. We anticipate that MLVU will advance the research of LVU by providing a comprehensive and in-depth analysis of MLLMs. The code and dataset can be accessed from https://github.com/JUNJIE99/MLVU.
****

Recovering Dynamic 3D Sketches from Videos
Jaeah Lee, Changwoon Choi, Young Min Kim, Jaesik Park; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12423-12432
Understanding 3D motion from videos presents inherent challenges due to the diverse types of movement, ranging from rigid and deformable objects to articulated structures. To overcome this, we propose Liv3Stroke, a novel approach for abstracting objects in motion with deformable 3D strokes. The detailed movements of an object may be represented by unstructured motion vectors or a set of motion primitives using a pre-defined articulation from a template model. Just as a free-hand sketch can intuitively visualize scenes or intentions with a sparse set of lines, we utilize a set of parametric 3D curves to capture a set of spatially smooth motion elements for general objects with unknown structures. We first extract noisy, 3D point cloud motion guidance from video frames using semantic features, and our approach deforms a set of curves to abstract essential motion features as a set of explicit 3D representations. Such abstraction enables an understanding of prominent components of motions while maintaining robustness to environmental factors. Our approach allows direct analysis of 3D object movements from video, tackling the uncertainty that typically occurs when translating real-world motion into recorded footage. The project page is accessible via: https://jaeah.me/liv3stroke_web.
****

IM-Zero: Instance-level Motion Controllable Video Generation in a Zero-shot Manner
Yuyang Huang, Yabo Chen, Li Ding, Xiaopeng Zhang, Wenrui Dai, Junni Zou, Hongkai Xiong, Qi Tian; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7265-7275
Controllability of video generation has been recently concerned in addition to the quality of generated videos. The main challenge to controllable video generation is to synthesize videos based on user-specified instance spatial locations and movement trajectories. However, existing methods suffer from a dilemma between the resource consumption, generation quality, and user controllability. As an efficient alternative to prohibitive training-based video generation, existing zero-shot video generation methods cannot generate high-quality and motion-consistent videos under the control of layouts and movement trajectories. In this paper, we propose a novel zero-shot method named IM-Zero that ameliorates instance-level motion controllable video generation with enhanced control accuracy, motion consistency, and richness of details to address this problem. Specifically, we first present a motion generation stage that extracts motion and textural guidance from keyframe candidates from pre-trained grounded text-to-image model to generate the desired coarse motion video. Subsequently, we develop a video refinement stage that injects the motion priors of pre-trained text-to-video models and detail priors of pre-trained text-to-image models into the latents of coarse motion videos to further enhance video motion consistency and richness of details. To our best knowledge, IM-Zero is the first to simultaneously achieve high-quality video generation and allow to control both layouts and movement trajectories in a zero-shot manner. Extensive experiments demonstrate that IM-Zero outperforms existing methods in terms of video quality, inter-frame consistency, and the alignment of location and trajectory. Furthermore, compared with existing methods, IM-Zero enjoys extra advantages of versatility in video generation, including motion control of subparts within instances, finer control of specifying instance shapes via masks, and more difficult tasks of motion transfer for customizing

fine-grained motion patterns through reference videos and high-quality text-to-video generation.
****

EigenGS Representation: From Eigenspace to Gaussian Image Space
Lo-Wei Tai, Ching-En Li, Cheng-Lin Chen, Chih-Jung Tsai, Hwann-Tzong Chen, Tyng-Luh Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13487-13496
Principal Component Analysis (PCA), a classical dimensionality reduction technique, and 2D Gaussian representation, an adaptation of 3D Gaussian Splatting for image representation, offer distinct approaches to modeling visual data. We present EigenGS, a novel method that bridges these paradigms through an efficient transformation pipeline connecting eigenspace and image-space Gaussian representations. Our approach enables instant initialization of Gaussian parameters for new images without requiring per-image optimization from scratch, dramatically accelerating convergence. EigenGS introduces a frequency-aware learning mechanism that encourages Gaussians to adapt to different scales, effectively modeling varied spatial frequencies and preventing artifacts in high-resolution reconstruction. Extensive experiments demonstrate that EigenGS not only achieves superior reconstruction quality compared to direct 2D Gaussian fitting but also reduces the necessary parameter count and training time. The results highlight EigenGS's effectiveness and generalization ability across images with varying resolutions and diverse categories, making Gaussian-based image representation both high-quality and viable for real-time applications.
****

Link-based Contrastive Learning for One-Shot Unsupervised Domain Adaptation
Yue Zhang, Mingyue Bin, Yuyang Zhang, Zhongyuan Wang, Zhen Han, Chao Liang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4916-4926
Unsupervised domain adaptation (UDA) aims to learn discriminative features from a labeled source domain by supervised learning and to transfer the knowledge to an unlabeled target domain via distribution alignment. However, in some real-world scenarios, e.g., public safety or access control, it's difficult to obtain sufficient source domain data, which hinders the application of existing UDA methods. To this end, this paper investigates a realistic but rarely studied problem called one-shot unsupervised domain adaptation (OSUDA), where there is only one example per category in the source domain and abundant unlabeled samples in the target domain. Compared with UDA, OSUDA faces dual challenges in both feature learning and domain alignment due to the lack of sufficient source data. To address these challenges, we propose a simple but effective link-based contrastive learning (LCL) method for OSUDA. On the one hand, with the help of in-domain links that indicate whether two samples are from the same cluster, LCL can learn discriminative features with abundant unlabeled target data. On the other hand, by constructing cross-domain links that show whether two clusters are bidirectionally matched, LCL can realize accurate domain alignment with only one source sample per category. Extensive experiments conducted on 4 public domain adaptation benchmarks, including VisDA-2017, Office-31, Office-Home, and DomainNet, demonstrate the effectiveness of the proposed LCL under the OSUDA setting. In addition, we build a realistic OSUDA surveillance video face recognition dataset, where LCL consistently improves the recognition accuracy across various face recognition methods.
****

SmartCLIP: Modular Vision-language Alignment with Identification Guarantees
Shaoan Xie, Lingjing Lingjing, Yujia Zheng, Yu Yao, Zeyu Tang, Eric P. Xing, Guangyi Chen, Kun Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29780-29790
Contrastive Language-Image Pre-training (CLIP) \citep radford2021learning has emerged as a pivotal model in computer vision and multimodal learning, achieving state-of-the-art performance at aligning visual and textual representations through contrastive learning. However, CLIP struggles with potential information misalignment in many image-text datasets and suffers from entangled representation.

On the one hand, short captions for a single image in datasets like MSCOCO may describe disjoint regions in the image, leaving the model uncertain about which visual features to retain or disregard. On the other hand, directly aligning long captions with images can lead to the retention of entangled details, preventing the model from learning disentangled, atomic concepts -- ultimately limiting its generalization on certain downstream tasks involving short prompts. In this paper, we establish theoretical conditions that enable flexible alignment between textual and visual representations across varying levels of granularity. Specifically, our framework ensures that a model can not only preserve cross-modal semantic information in its entirety but also disentangle visual representations to capture fine-grained textual concepts. Building on this foundation, we introduce \ours, a novel approach that identifies and aligns the most relevant visual and textual representations in a modular manner. Superior performance across various tasks demonstrates its capability to handle information misalignment and supports our identification theory. The code is available at https://github.com/Mid-Push/SmartCLIP.

****

UniMamba: Unified Spatial-Channel Representation Learning with Group-Efficient Mamba for LiDAR-based 3D Object Detection

Xin Jin, Haisheng Su, Kai Liu, Cong Ma, Wei Wu, Fei HUI, Junchi Yan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1407-1417

Recent advances in LiDAR 3D detection have demonstrated the effectiveness of Transformer-based frameworks in capturing the global dependencies from point cloud spaces, which serialize the 3D voxels into the flattened 1D sequence for iterative self-attention. However, the spatial structure of 3D voxels will be inevitably destroyed during the serialization process. Besides, due to the considerable number of 3D voxels and quadratic complexity of Transformers, multiple sequences are grouped before feeding to Transformers, leading to a limited receptive field. Inspired by the impressive performance of State Space Models (SSM), in this paper, we propose a novel Unified Mamba (UniMamba), which seamlessly integrates the merits of 3D convolution and SSM in a concise multi-head manner, aiming to perform "local and global" spatial context aggregation efficiently and simultaneously. Specifically, a UniMamba block is designed which mainly consists of spatial locality modeling, complementary Z-order serialization and local-global sequential aggregator. The spatial locality modeling module integrates 3D submanifold convolution to capture the dynamic spatial position embedding before serialization. Then the efficient Z-order curve is adopted for serialization both horizontally and vertically. Furthermore, the local-global sequential aggregator adopts the channel grouping strategy to efficiently encode both "local and global" spatial inter-dependencies using multi-head SSM. Additionally, an encoder-decoder architecture with stacked UniMamba blocks is formed to facilitate multi-scale spatial learning hierarchically. Extensive experiments are conducted on three popular datasets: nuScenes, Waymo and Argoverse 2. Particularly, our UniMamba achieves 70.2 mAP on the nuScenes dataset.

****

MaSS13K: A Matting-level Semantic Segmentation Benchmark

Chenxi Xie, Minghan Li, Hui Zeng, Jun Luo, Lei Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14046-14056

High-resolution semantic segmentation is essential for applications such as image editing, bokeh imaging, AR/VR, etc. Unfortunately, existing datasets often have limited resolution and lack precise mask details and boundaries. In this work, we build a large-scale, matting-level semantic segmentation dataset, named MaSS13K, which consists of 13,348 real-world images, all at 4K resolution. MaSS13K provides high-quality mask annotations of a number of objects, which are categorized into seven categories: human, vegetation, ground, sky, water, building, and others. MaSS13K features precise masks, with an average mask complexity 20-50 times higher than existing semantic segmentation datasets. We consequently present a method specifically designed for high-resolution semantic segmentation, namely MaSSFormer, which employs an efficient pixel decoder that aggregates high-leve

l semantic features and low-level texture features across three stages, aiming t
o produce high-resolution masks with minimal computational cost. Finally, we pro
pose a new learning paradigm, which integrates the high-quality masks of the sev
en given categories with pseudo labels from new classes, enabling MaSSFormer to
transfer its accurate segmentation capability to other classes of objects. Our p
roposed MaSSFormer is comprehensively evaluated on the MaSS13K benchmark togethe
r with 14 representative segmentation models. We expect that our meticulously a
nnotated MaSS13K dataset and the MaSSFormer model can facilitate the research of
 high-resolution and high-quality semantic segmentation. Datasets and codes can
be found at https://github.com/xiechenxi99/MaSS13K.
****

Rethinking the Adversarial Robustness of Multi-Exit Neural Networks in an Attack
-Defense Game
Keyizhi Xu, Chi Zhang, Zhan Chen, Zhongyuan Wang, Chunxia Xiao, Chao Liang; Proc
eedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025,
pp. 10265-10274
Multi-exit neural networks represent a promising approach to enhancing model inf
erence efficiency, yet like common neural networks, they suffer from significant
ly reduced robustness against adversarial attacks. While some defense methods ha
ve been raised to strengthen the adversarial robustness of multi-exit neural net
works, we identify a long-neglected flaw in the evaluation of previous studies:
simply using a fixed set of exits for attack may lead to an overestimation of th
eir defense capacity. Based on this finding, our work explores the following thr
ee key aspects in the adversarial robustness of multi-exit neural networks: (1)
we discover that a mismatch of the network exits used by the attacker and defend
er is responsible for the overestimated robustness of previous defense methods;
(2) by finding the best strategy in a two-player zero-sum game, we propose AIMER
 as an improved evaluation scheme to measure the intrinsic robustness of multi-e
xit neural networks; (3) going further, we introduce NEED defense method under t
he evaluation of AIMER that can optimize the defender's strategy by finding a Na
sh equilibrium of the game. Experiments over 3 datasets, 7 architectures, 7 atta
cks and 4 baselines show that AIMER evaluates the robustness 13.52% lower than p
revious methods under AutoAttack, while the robust performance of NEED surpasses
 single-exit networks of the same backbones by 5.58% maximally.
****

Enhancing Testing-Time Robustness for Trusted Multi-View Classification in the W
ild
Wei Liu, Yufei Chen, Xiaodong Yue; Proceedings of the Computer Vision and Patter
n Recognition Conference (CVPR), 2025, pp. 15508-15517
Trusted multi-view classification (TMVC) addresses variations in data quality by
 evaluating the reliability of each view based on prediction uncertainty at the
evidence level, reducing the impact of low-quality views commonly encountered in
 real-world scenarios. However, existing TMVC methods often struggle to maintain
 robustness during testing, particularly when integrating noisy or corrupted vie
ws. This limitation arises because the evidence collected by TMVC may be unreli
able, frequently providing incorrect information due to complex view distributio
ns and optimization challenges, ultimately leading to classification performance
 degradation. To enhance the robustness of TMVC methods in real-world conditions
, we propose a generalized evidence filtering mechanism that is compatible with
various fusion strategies commonly used in TMVC, including Belief Constraint Fus
ion, Aleatory Cumulative Belief Fusion, and Averaging Belief Fusion. Specificall
y, we frame the identification of unreliable evidence as a multiple testing prob
lem and introduce p-values to control the risk of false identification. By selec
tively down-weighting unreliable evidence during testing, our mechanism ensures
robust fusion and mitigates performance degradation. Both theoretical guarantees
 and empirical results demonstrate significant improvements in the classificatio
n performance of TMVC methods, supporting their reliable application in challeng
ing, real-world environments.
****

Q-DiT: Accurate Post-Training Quantization for Diffusion Transformers

Lei Chen, Yuan Meng, Chen Tang, Xinzhu Ma, Jingyan Jiang, Xin Wang, Zhi Wang, We
nwu Zhu; Proceedings of the Computer Vision and Pattern Recognition Conference (
CVPR), 2025, pp. 28306-28315
Recent advancements in diffusion models, particularly the architectural transfor
mation from UNet-based models to Diffusion Transformers (DiTs), significantly im
prove the quality and scalability of image and video generation. However, despit
e their impressive capabilities, the substantial computational costs of these la
rge-scale models pose significant challenges for real-world deployment. Post-Tra
ining Quantization (PTQ) emerges as a promising solution, enabling model compres
sion and accelerated inference for pretrained models, without the costly retrain
ing. However, research on DiT quantization remains sparse, and existing PTQ fram
eworks, primarily designed for traditional diffusion models, tend to suffer from
 biased quantization, leading to notable performance degradation. In this work,
we identify that DiTs typically exhibit significant spatial variance in both wei
ghts and activations, along with temporal variance in activations. To address th
ese issues, we propose Q-DiT, a novel approach that seamlessly integrates two ke
y techniques: automatic quantization granularity allocation to handle the signif
icant variance of weights and activations across input channels, and sample-wise
 dynamic activation quantization to adaptively capture activation changes across
 both timesteps and samples. Extensive experiments conducted on ImageNet and VBe
nch demonstrate the effectiveness of the proposed Q-DiT. Specifically, when quan
tizing DiT-XL/2 to W6A8 on ImageNet (256 x256), Q-DiT achieves a remarkable redu
ction in FID by 1.09 compared to the baseline. Under the more challenging W4A8 s
etting, it maintains high fidelity in image and video generation, establishing a
 new benchmark for efficient, high-quality quantization in DiTs.
****

ROD-MLLM: Towards More Reliable Object Detection in Multimodal Large Language Mo
dels
Heng Yin, Yuqiang Ren, Ke Yan, Shouhong Ding, Yongtao Hao; Proceedings of the Co
mputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14358-14368
Multimodal large language models (MLLMs) have demonstrated strong language under
standing and generation capabilities, excelling in visual tasks like referring a
nd grounding. However, due to task type limitations and dataset scarcity, existi
ng MLLMs only ground objects present in images and cannot reject non-existent ob
jects effectively, resulting in unreliable predictions. In this paper, we introd
uce ROD-MLLM, a novel MLLM for Reliable Object Detection using free-form languag
e. We propose a query-based localization mechanism to extract low-level object f
eatures. By aligning global and object-level visual information with text space,
 we leverage the large language model (LLM) for high-level comprehension and fin
al localization decisions, overcoming the language understanding limitations of
normal detectors. To enhance language-based object detection, we design an autom
ated data annotation pipeline and construct the dataset ROD. This pipeline uses
the referring capabilities of existing MLLMs and chain-of-thought techniques to
generate diverse expressions corresponding to zero or multiple objects, addressi
ng the shortage of training data. Experiments across various tasks, including re
ferring, grounding, and language-based object detection, show that ROD-MLLM achi
eves state-of the-art performance among MLLMs. Notably, in language-based object
 detection, our model achieves +13.7 AP improvement on D3 benchmark over existin
g MLLMs and surpasses most specialized detection models, especially in scenarios
 requiring complex language understanding.
****

RoboGround: Robotic Manipulation with Grounded Vision-Language Priors
Haifeng Huang, Xinyi Chen, Yilun Chen, Hao Li, Xiaoshen Han, Zehan Wang, Tai Wan
g, Jiangmiao Pang, Zhou Zhao; Proceedings of the Computer Vision and Pattern Rec
ognition Conference (CVPR), 2025, pp. 22540-22550
Recent advancements in robotic manipulation have highlighted the potential of in
termediate representations for improving policy generalization. In this work, we
 explore grounding masks as an effective intermediate representation, balancing
two key advantages: (1) effective spatial guidance that specifies target objects
 and placement areas while also conveying information about object shape and siz

e, and (2) broad generalization potential driven by large-scale vision-language models pretrained on diverse grounding datasets. We introduce \method, a grounding-aware robotic manipulation policy that leverages grounding masks as an intermediate representation to guide policy networks in object manipulation tasks. To further explore and enhance generalization, we propose an automated pipeline for generating large-scale, simulated data with a diverse set of objects and instructions. Extensive experiments show the value of our dataset and the effectiveness of grounding masks as intermediate guidance, significantly enhancing the generalization abilities of robot policies.
****

VideoGuide: Improving Video Diffusion Models without Training Through a Teacher's Guide
Dohun Lee, Bryan Sangwoo Kim, Geon Yeong Park, Jong Chul Ye; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2599-2608
Text-to-image (T2I) diffusion models have revolutionized visual content creation, but extending these capabilities to text-to-video (T2V) generation remains a challenge, particularly in preserving temporal consistency. Existing methods that aim to improve consistency often cause trade-offs such as reduced imaging quality and impractical computational time. To address these issues we introduce VideoGuide, a novel framework that enhances the temporal consistency of pretrained T2V models without the need for additional training or fine-tuning. Instead, VideoGuide leverages any pretrained video diffusion model (VDM) or itself as a guide during the early stages of inference, improving temporal quality by interpolating the guiding model's denoised samples into the sampling model's denoising process. The proposed method brings about significant improvement in temporal consistency and image fidelity, providing a cost-effective and practical solution that synergizes the strengths of various video diffusion models. Furthermore, we demonstrate prior distillation, revealing that base models can achieve enhanced text coherence by utilizing the superior data prior of the guiding model through the proposed method. Project Page: https://dohunlee1.github.io/videoguide.github.io/
****

Improving Transferable Targeted Attacks with Feature Tuning Mixup
Kaisheng Liang, Xuelong Dai, Yanjie Li, Dong Wang, Bin Xiao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25802-25811
Deep neural networks (DNNs) exhibit vulnerability to adversarial examples that can transfer across different DNN models. A particularly challenging problem is developing transferable targeted attacks that can mislead DNN models into predicting specific target classes. While various methods have been proposed to enhance attack transferability, they often incur substantial computational costs while yielding limited improvements. Recent clean feature mixup methods use random clean features to perturb the feature space but lack optimization for disrupting adversarial examples, overlooking the advantages of attack-specific perturbations. In this paper, we propose Feature Tuning Mixup (FTM), a novel method that enhances targeted attack transferability by combining both random and optimized noises in the feature space. FTM introduces learnable feature perturbations and employs an efficient stochastic update strategy for optimization. These learnable perturbations facilitate the generation of more robust adversarial examples with improved transferability. We further demonstrate that attack performance can be enhanced through an ensemble of multiple FTM-perturbed surrogate models. Extensive experiments on the ImageNet-compatible dataset across various DNN models demonstrate that our method achieves significant improvements over state-of-the-art methods while maintaining low computational cost.
****

OmniStereo: Real-time Omnidireactional Depth Estimation with Multiview Fisheye Cameras
Jiaxi Deng, Yushen Wang, Haitao Meng, Zuoxun Hou, Yi Chang, Gang Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1003-1012
Fast and reliable omnidirectional 3D sensing is essential to many applications s

uch as autonomous driving, robotics and drone navigation. While many well-recogn ized methods have been developed to produce high-quality omnidirectional 3D info rmation, they are too slow for real-time computation, limiting their feasibility in practical applications. Motivated by these shortcomings, we propose an effic ient omnidirectional depth sensing framework, called OmniStereo, which generates high-quality 3D information in real-time. Unlike prior works, OmniStereo employ s Cassini projection to simplify the photometric matching and introduces a light weight stereo matching network to minimize computational overhead. Additionally, OmniStereo proposes a novel fusion method to handle depth discontinuities and i nvalid pixels complemented by a refinement module to reduce mapping-introduced e rrors and recover fine details. As a result, OmniStereo achieves state-of-the-ar t (SOTA) accuracy, surpassing the second-best method over 32% in MAE, while main taining real-time efficiency. It operates more than 16.5xfaster than the second- best method in accuracy on TITAN RTX, achieving 12.3 FPS on embedded device Jets on AGX Orin, underscoring its suitability for real-world deployment. The code is available at https://github.com/DengJiaxi1/OmniStereo.

****

DroneSplat: 3D Gaussian Splatting for Robust 3D Reconstruction from In-the-Wild Drone Imagery

Jiadong Tang, Yu Gao, Dianyi Yang, Liqi Yan, Yufeng Yue, Yi Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 833-84 3

Drones have become essential tools for reconstructing wild scenes due to their o utstanding maneuverability. Recent advances in radiance field methods have achie ved remarkable rendering quality, providing a new avenue for 3D reconstruction f rom drone imagery. However, dynamic distractors in wild environments challenge the static scene assumption in radiance fields, while limited view constraints h inder the accurate capture of underlying scene geometry. To address these challe nges, we introduce DroneSplat, a novel framework designed for robust 3D reconstr uction from in-the-wild drone imagery. Our method adaptively adjusts masking thr esholds by integrating local-global segmentation heuristics with statistical app roaches, enabling precise identification and elimination of dynamic distractors in static scenes. We enhance 3D Gaussian Splatting with multi-view stereo predic tions and a voxel-guided optimization strategy, supporting high-quality renderin g under limited view constraints. For comprehensive evaluation, we provide a dro ne-captured 3D reconstruction dataset encompassing both dynamic and static scene s. Extensive experiments demonstrate that DroneSplat outperforms both 3DGS and N eRF baselines in handling in-the-wild drone imagery.

****

SDGOCC: Semantic and Depth-Guided Bird's-Eye View Transformation for 3D Multimod al Occupancy Prediction

ZaiPeng Duan, ChenXu Dang, Xuzhong Hu, Pei An, Junfeng Ding, Jie Zhan, YunBiao X u, Jie Ma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6751-6760

Multimodal 3D occupancy prediction has garnered significant attention for its po tential in autonomous driving. However, most existing approaches are single-moda lity: camera-based methods lack depth information, while LiDAR-based methods str uggle with occlusions. Current lightweight methods primarily rely on the Lift-Sp lat-Shoot (LSS) pipeline, which suffers from inaccurate depth estimation and fai ls to fully exploit the geometric and semantic information of 3D LiDAR points. T herefore, we propose a novel multimodal occupancy prediction network called SDG- OCC, which incorporates a joint semantic and depth-guided view transformation co upled with a fusion-to-occupancy-driven active distillation. The enhanced view t ransformation constructs accurate depth distributions by integrating pixel seman tics and co-point depth through diffusion and bilinear discretization. The fusio n-to-occupancy-driven active distillation extracts rich semantic information fro m multimodal data and selectively transfers knowledge to image features based on LiDAR-identified regions. Finally, for optimal performance, we introduce SDG-Fu sion, which uses fusion alone, and SDG-KL, which integrates both fusion and dist illation for faster inference. Our method achieves state-of-the-art (SOTA) perfo

rmance with real-time processing on the Occ3D-nuScenes dataset and shows compara ble performance on the more challenging SurroundOcc-nuScenes dataset, demonstrat ing its effectiveness and robustness. The code will be released at https://githu b.com/DzpLab/SDGOCC.
****

DrivingSphere: Building a High-fidelity 4D World for Closed-loop Simulation
Tianyi Yan, Dongming Wu, Wencheng Han, Junpeng Jiang, Xia Zhou, Kun Zhan, Cheng- zhong Xu, Jianbing Shen; Proceedings of the Computer Vision and Pattern Recognit ion Conference (CVPR), 2025, pp. 27531-27541
Autonomous driving evaluation requires simulation environments that closely repl icate actual road conditions, including real-world sensory data and responsive f eedback loops. However, many existing simulations need to predict waypoints alon g fixed routes on public datasets or synthetic photorealistic data, i.e., open-l oop simulation usually lacks the ability to assess dynamic decision-making. Whil e the recent efforts of closed-loop simulation offer feedback-driven environment s, they cannot process visual sensor inputs or produce outputs that differ from real-world data. To address these challenges, we propose DrivingSphere, a realis tic and closed-loop simulation framework. Its core idea is to build 4D world rep resentation and generate real-life and controllable driving scenarios. In specif ic, our framework includes a Dynamic Environment Composition module that constru cts a detailed 4D driving world with a format of occupancy equipping with static backgrounds and dynamic objects, and a Visual Scene Synthesis module that trans forms this data into high-fidelity, multi-view video outputs, ensuring spatial a nd temporal consistency. By providing a dynamic and realistic simulation environ ment, DrivingSphere enables comprehensive testing and validation of autonomous d riving algorithms, ultimately advancing the development of more reliable autonom ous cars.The benchmark will be publicly released.
****

nnWNet: Rethinking the Use of Transformers in Biomedical Image Segmentation and Calling for a Unified Evaluation Benchmark
Yanfeng Zhou, Lingrui Li, Le Lu, Minfeng Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20852-20862
Semantic segmentation is a crucial prerequisite in clinical applications and com puter-aided diagnosis. With the development of deep neural networks, biomedical image segmentation has achieved remarkable success. Encoder-decoder architecture s that integrate convolutions and transformers are gaining attention for their p otential to capture both global and local features. However, current designs fac e the contradiction that these two features cannot be continuously transmitted. In addition, some models lack a unified and standardized evaluation benchmark, l eading to significant discrepancies in the experimental setup. In this study, we review and summarize these architectures and analyze their contradictions in de sign. We modify UNet and propose WNet to combine transformers and convolutions, addressing the transmission issue effectively. WNet captures long-range dependen cies and local details simultaneously while ensuring their continuous transmissi on and multi-scale fusion. We integrate WNet into the nnUNet framework for unifi ed benchmarking. Our model achieves state-of-the-art performance in biomedical i mage segmentation. Extensive experiments demonstrate their effectiveness on four 2D datasets (DRIVE, ISIC-2017, Kvasir-SEG, and CREMI) and four 3D datasets (Par se2022, AMOS22, BTCV, and ImageCAS). The code is available at https://github.com /Yanfeng-Zhou/nnWNet.
****

Efficient Video Face Enhancement with Enhanced Spatial-Temporal Consistency
Yutong Wang, Jiajie Teng, Jiajiong Cao, Yuming Li, Chenguang Ma, Hongteng Xu, Di xin Luo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2183-2193
As a very common type of video, face videos often appear in movies, talk shows, live broadcasts, and other scenes. Real-world online videos are often plagued by degradations such as blurring and quantization noise, due to the high compressi on ratio caused by high communication costs and limited transmission bandwidth. These degradations have a particularly serious impact on face videos because the

human visual system is highly sensitive to facial details.Despite the significant advancement in video face enhancement, current methods still suffer from i) long processing time and ii) inconsistent spatial-temporal visual effects (e.g., flickering). This study proposes a novel and efficient blind video face enhancement method to overcome the above two challenges, restoring high-quality videos from their compressed low-quality versions with an effective de-flickering mechanism. In particular, the proposed method develops upon a 3D-VQGAN backbone associated with spatial-temporal codebooks recording high-quality portrait features and residual-based temporal information. We develop a two-stage learning framework for the model. In Stage I, we learn the model with a regularizer mitigating the codebook collapse problem.In Stage II, we learn two transformers to lookup code from the codebooks and further update the encoder of low-quality videos.Experiments conducted on the VFHQ-Test dataset demonstrate that our method surpasses the current state-of-the-art blind face video restoration and de-flickering methods on both efficiency and effectiveness.

****

VELOCITI: Benchmarking Video-Language Compositional Reasoning with Strict Entailment

Darshana Saravanan, Varun Gupta, Darshan Singh, Zeeshan Khan, Vineet Gandhi, Makarand Tapaswi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18914-18924

A fundamental aspect of compositional reasoning in a video is associating people and their actions across time. Recent years have seen great progress in general-purpose vision/video models and a move towards long-video understanding. While exciting, we take a step back and ask: are today's models good at compositional reasoning on short videos? To this end, we introduce VELOCITI, a benchmark to study Video-LLMs by disentangling and assessing the comprehension of agents, actions, and their associations across multiple events. We adopt the Video-Language Entailment setup and propose StrictVLE that requires correct classification (rather than ranking) of the positive and negative caption. We evaluate several models and observe that even the best, LLaVA-OneVision (44.5%) and Gemini-1.5-Pro (49.3%), are far from human accuracy at 93.0%. Results show that action understanding lags behind agents, and negative captions created using entities appearing in the video perform worse than those obtained from pure text manipulation. We also present challenges with ClassicVLE and multiple-choice (MC) evaluation, strengthening our preference for StrictVLE. Finally, we validate that our benchmark requires visual inputs of multiple frames making it ideal to study video-language compositional reasoning.

****

Seeing is Not Believing: Adversarial Natural Object Optimization for Hard-Label 3D Scene Attacks

Daizong Liu, Wei Hu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11886-11897

Deep learning models for 3D data have shown to be vulnerable to adversarial attacks, which have received increasing attention in various safety-critical applications such as autonomous driving and robotic navigation. Existing 3D attackers mainly put effort into attacking the simple 3D classification model by perturbing point cloud objects in the white/black-box setting. However, real-world 3D applications focus on tackling more complicated scene-based data while sharing no information about the model parameters and logits with users. Therefore, directly applying previous naive 3D attack methods to these applications does not work. To this end, this paper attempts to address the challenging hard-label 3D scene attack with access only to the input/output of the 3D models. To make the attack effective and stealthy, we propose to generate universal adversarial objects, which will mislead scene-aware 3D models to predict attacker-chosen labels whenever these objects are placed on any scene input. Specifically, we inject an imperceptible object trigger with further perturbations into all scenes and learn to mislead their reasoning by only querying the 3D model. We start by initializing the trigger pattern with a realistic object and searching for an appropriate location to place it naturally in the scene data. Then, we design a novel weighted g

radient estimation strategy to perturb the object trigger with additive slight n
oise to make them adversarial in an iterative optimization procedure. Extensive
experiments demonstrate that our attack can achieve superior performance on seve
n 3D models and three scene-based datasets, with satisfactory adversarial imperc
eptibility and strong resistance to defense methods.
****

IDProtector: An Adversarial Noise Encoder to Protect Against ID-Preserving Image
 Generation
Yiren Song, Pei Yang, Hai Ci, Mike Zheng Shou; Proceedings of the Computer Visio
n and Pattern Recognition Conference (CVPR), 2025, pp. 3019-3028
Recently, zero-shot methods like InstantID have revolutionized identity-preservi
ng generation. Unlike multi-image finetuning approaches such as DreamBooth, thes
e zero-shot methods leverage powerful facial encoders to extract identity inform
ation from a single portrait photo, enabling efficient identity-preserving gener
ation through a single inference pass. However, this convenience introduces new
threats to the facial identity protection. This paper aims to safeguard portrait
 photos from unauthorized encoder-based customization. We introduce IDProtector,
 an adversarial noise encoder that applies imperceptible adversarial noise to po
rtrait photos in a single forward pass. Our approach offers universal protection
 for portraits against multiple state-of-the-art encoder-based methods, includin
g InstantID, IP-Adapter, and PhotoMaker, while ensuring robustness to common ima
ge transformations such as JPEG compression, resizing, and affine transformation
s. Experiments across diverse portrait datasets and generative models reveal tha
t IDProtector generalizes effectively to unseen data and even closed-source prop
rietary models.
****

HuPerFlow: A Comprehensive Benchmark for Human vs. Machine Motion Estimation Com
parison
Yung-Hao Yang, Zitang Sun, Taiki Fukiage, Shin'ya Nishida; Proceedings of the Co
mputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22799-22808
As AI models are increasingly integrated into applications involving human inter
action, understanding the alignment between human perception and machine vision
has become essential. One example is the estimation of visual motion (optical fl
ow) in dynamic applications such as driving assistance. While there are numerous
 optical flow datasets and benchmarks with ground truth information, human-perce
ived flow in natural scenes remains underexplored. We introduce HuPerFlow--a ben
chmark for human-perceived flow, measured at 2,400 locations across ten optical
flow datasets, with 38,400 response vectors collected through online psychophys
ical experiments. Our data demonstrate that human-perceived flow aligns with gro
und truth in spatiotemporally smooth locations while also showing systematic err
ors influenced by various environmental properties. Additionally, we evaluated s
everal optical flow algorithms against human-perceived flow, uncovering both sim
ilarities and unique aspects of human perception in complex natural scenes. HuPe
rFlow is the first large-scale human-perceived flow benchmark for alignment betw
een computer vision models and human perception, as well as for scientific explo
ration of human motion perception in natural scenes. The HuPerFlow benchmark wil
l be available online upon acceptance.
****

LoTUS: Large-Scale Machine Unlearning with a Taste of Uncertainty
Christoforos N. Spartalis, Theodoros Semertzidis, Efstratios Gavves, Petros Dara
s; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR),
 2025, pp. 10046-10055
We present LoTUS, a novel Machine Unlearning (MU) method that eliminates the inf
luence of training samples from pre-trained models, avoiding retraining from scr
atch. LoTUS smooths the prediction probabilities of the model up to an informati
on-theoretic bound, mitigating its over-confidence stemming from data memorizati
on. We evaluate LoTUS on Transformer and ResNet18 models against eight baselines
 across five public datasets. Beyond established MU benchmarks, we evaluate unle
arning on ImageNet1k, a large-scale dataset, where retraining is impractical, si
mulating real-world conditions. Moreover, we introduce the novel Retrain-Free Je

nsen-Shannon Divergence (RF-JSD) metric to enable evaluation under real-world co nditions. The experimental results show that LoTUS outperforms state-of-the-art methods in terms of both efficiency and effectiveness. Code: https://github.com/cspartalis/LoTUS

****

SleeperMark: Towards Robust Watermark against Fine-Tuning Text-to-image Diffusio n Models

Zilan Wang, Junfeng Guo, Jiacheng Zhu, Yiming Li, Heng Huang, Muhao Chen, Zhengz hong Tu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8213-8224

Recent advances in large-scale text-to-image (T2I) diffusion models have enabled a variety of downstream applications. As T2I models require extensive resources for training, they constitute highly valued intellectual property (IP) for thei r legitimate owners, yet making them incentive targets for unauthorized fine-tun ing by adversaries seeking to leverage these models for customized, usually prof itable applications. Existing IP protection methods for diffusion models general ly involve embedding watermark patterns and then verifying ownership through gen erated outputs examination, or inspecting the model's feature space. However, th ese techniques are inherently ineffective in practical scenarios when the waterm arked model undergoes fine-tuning, and the feature space is inaccessible during verification (i.e., black-box setting). The model is prone to forgetting the pre viously learned watermark knowledge when it adapts to a new task. To address thi s challenge, we propose SleeperMark, a novel framework designed to embed resilie nt watermarks into T2I diffusion models. SleeperMark explicitly guides the model to disentangle the watermark information from the semantic concepts it learns, allowing the model to retain the embedded watermark while continuing to be adapt ed to new downstream tasks. Our extensive experiments demonstrate the effectiven ess of SleeperMark across various types of diffusion models, including latent di ffusion models (e.g., Stable Diffusion) and pixel diffusion models (e.g., DeepFl oyd-IF), showing robustness against downstream fine-tuning and various attacks a t both the image and model levels, with minimal impact on the model's generative capability.

****

Lessons and Insights from a Unifying Study of Parameter-Efficient Fine-Tuning (P EFT) in Visual Recognition

Zheda Mai, Ping Zhang, Cheng-Hao Tu, Hong-You Chen, Quang-Huy Nguyen, Li Zhang, Wei-Lun Chao; Proceedings of the Computer Vision and Pattern Recognition Confere nce (CVPR), 2025, pp. 14845-14857

Parameter-efficient fine-tuning (PEFT) has attracted significant attention due t o the growth of pre-trained model sizes and the need to fine-tune (FT) them for superior downstream performance. Despite a surge in new PEFT methods, a systemat ic study to understand their performance and suitable application scenarios is l acking, leaving questions like "when to apply PEFT" and "which method to use" la rgely unanswered, especially in visual recognition. In this paper, we conduct a unifying empirical study of representative PEFT methods with Vision Transformers . We systematically tune their hyperparameters to fairly compare their accuracy on downstream tasks. Our study offers a practical user guide and unveils several new insights. First, if tuned carefully, different PEFT methods achieve similar accuracy in the low-shot benchmark VTAB-1K. This includes simple approaches lik e FT the bias terms that were reported inferior. Second, despite similar accurac y, we find that PEFT methods make different mistakes and high-confidence predict ions, likely due to their different inductive biases. Such an inconsistency (or complementarity) opens up the opportunity for ensemble methods, and we make prel iminary attempts at this. Third, going beyond the commonly used low-shot tasks, we find that PEFT is also useful in many-shot regimes, achieving comparable or b etter accuracy than full FT while using significantly fewer parameters. Lastly, we investigate PEFT's ability to preserve a pre-trained model's robustness to di stribution shifts (e.g., CLIP). Perhaps not surprisingly, PEFT approaches outper form full FT alone. However, with weight-space ensembles, full FT can better ba lance target distribution and distribution shift performance, suggesting a futur

e research direction for robust PEFT. The code is available at https://github.com/OSU-MLB/ViT_PEFT_Vision.
****

Pippo: High-Resolution Multi-View Humans from a Single Image
Yash Kant, Ethan Weber, Jin Kyu Kim, Rawal Khirodkar, Su Zhaoen, Julieta Martinez, Igor Gilitschenski, Shunsuke Saito, Timur Bagautdinov; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16418-16429
We present Pippo, a generative model capable of producing 1K resolution dense turnaround videos of a person from a single casually clicked photo. Pippo is a multi-view diffusion transformer and does not require any additional inputs - e.g., a fitted parametric model or camera parameters of the input image. We pre-train Pippo on 3B human images without captions, and conduct multi-view mid-training and post-training on studio captured humans. During mid-training, to quickly absorb the studio dataset, we denoise several (up to 48) views at low-resolution, and encode target cameras coarsely using a shallow MLP. During post-training, we denoise fewer views at high-resolution and use pixel-aligned controls (e.g., Spatial anchor and Plucker rays) to enable 3D consistent generations. At inference, we propose an attention biasing technique that allows Pippo to simultaneously generate greater than 5 times as many views as seen during training. Finally, we also introduce an improved metric to evaluate 3D consistency of multi-view generations, and show that Pippo outperforms existing works on multi-view human generation from a single image.
****

H2ST: Hierarchical Two-Sample Tests for Continual Out-of-Distribution Detection
Yuhang Liu, Wenjie Zhao, Yunhui Guo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15413-15423
Task Incremental Learning (TIL) is a specialized form of Continual Learning (CL) in which a model incrementally learns from non-stationary data streams. Existing TIL methodologies operate under the closed-world assumption, presuming that incoming data remains in-distribution (ID). However, in an open-world setting, incoming samples may originate from out-of-distribution (OOD) sources, with their task identities inherently unknown. Continually detecting OOD samples presents several challenges for current OOD detection methods: reliance on model outputs leads to excessive dependence on model performance, selecting suitable thresholds is difficult, hindering real-world deployment, and binary ID/OOD classification fails to provide task-level identification. To address these issues, we propose a novel continual OOD detection method called the Hierarchical Two-sample Tests (H2ST). H2ST eliminates the need for threshold selection through hypothesis testing and utilizes feature maps to better exploit model capabilities without excessive dependence on model performance. The proposed hierarchical architecture enables task-level detection with superior performance and lower overhead compared to non-hierarchical classifier two-sample tests. Extensive experiments and analysis validate the effectiveness of H2ST in open-world TIL scenarios and its superiority to the existing methods. Code is available at https://github.com/YuhangLiuu/H2ST.
****

MetaWriter: Personalized Handwritten Text Recognition Using Meta-Learned Prompt Tuning
Wenhao Gu, Li Gu, Chingyee Yee Suen, Yang Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23494-23504
Recent advancements in handwritten text recognition (HTR) have enabled the effective conversion of handwritten text to digital formats. However, achieving robust recognition across diverse writing styles remains challenging. Traditional HTR methods lack writer-specific personalization at test time due to limitations in model architecture and training strategies. Existing attempts to bridge this gap, through gradient-based meta-learning, still require labeled examples and suffer from parameter-inefficient fine-tuning, leading to substantial computational and memory overhead. To overcome these challenges, we propose an efficient framework that formulates personalization as prompt tuning, incorporating an auxiliary image reconstruction task with a self-supervised loss to guide prompt adaptati

on with unlabeled test-time examples. To ensure self-supervised loss effectively minimizes text recognition error, we leverage meta-learning to learn the optimal initialization of the prompts. As a result, our method allows the model to efficiently capture unique writing styles by updating less than 1% of its parameters and eliminating the need for time-intensive annotation processes. We validate our approach on the RIMES and IAM Handwriting Database benchmarks, where it consistently outperforms previous state-of-the-art methods while using 20x fewer parameters. We believe this represents a significant advancement in personalized handwritten text recognition, paving the way for more reliable and practical deployment in resource-constrained scenarios.
****

Subnet-Aware Dynamic Supernet Training for Neural Architecture Search
Jeimin Jeon, Youngmin Oh, Junghyup Lee, Donghyeon Baek, Dohyung Kim, Chanho Eom, Bumsub Ham; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30137-30146
N-shot neural architecture search (NAS) exploits a supernet containing all candidate subnets for a given search space. The subnets are typically trained with a static training strategy (e.g., using the same learning rate (LR) scheduler and optimizer for all subnets). This, however, does not consider that individual subnets have distinct characteristics, leading to two problems: (1) The supernet training is biased towards the low-complexity subnets (unfairness); (2) the momentum update in the supernet is noisy (noisy momentum). We present a dynamic supernet training technique to address these problems by adjusting the training strategy adaptive to the subnets. Specifically, we introduce a complexity-aware LR scheduler (CaLR) that controls the decay ratio of LR adaptive to the complexities of subnets, which alleviates the unfairness problem. We also present a momentum separation technique (MS). It groups the subnets with similar structural characteristics and uses a separate momentum for each group, avoiding the noisy momentum problem. Our approach can be applicable to various N-shot NAS methods with marginal cost, while improving the search performance drastically. We validate the effectiveness of our approach on various search spaces (e.g., NAS-Bench-201, Mobilenet spaces) and datasets (e.g., CIFAR-10/100, ImageNet).
****

MoVE-KD: Knowledge Distillation for VLMs with Mixture of Visual Encoders
Jiajun Cao, Yuan Zhang, Tao Huang, Ming Lu, Qizhe Zhang, Ruichuan An, Ningning Ma, Shanghang Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19846-19856
Visual encoders are fundamental components in vision-language models (VLMs), each showcasing unique strengths derived from various pre-trained visual foundation models. To leverage the various capabilities of these encoders, recent studies incorporate multiple encoders within a single VLM, leading to a considerable increase in computational cost. In this paper, we present Mixture-of-Visual-Encoder Knowledge Distillation (MoVE-KD), a novel framework that distills the unique proficiencies of multiple vision encoders into a single, efficient encoder model. Specifically, to mitigate conflicts and retain the unique characteristics of each teacher encoder, we employ low-rank adaptation (LoRA) and mixture-of-experts (MoEs) to selectively activate specialized knowledge based on input features, enhancing both adaptability and efficiency. To regularize the KD process and enhance performance, we propose an attention-based distillation strategy that adaptively weighs the different encoders and emphasizes valuable visual tokens, reducing the burden of replicating comprehensive but distinct features from multiple teachers. Comprehensive experiments on popular VLMs, such as LLaVA and LLaVA-NeXT, validate the effectiveness of our method. Our code is available at:https://github.com/hey-cjj/MoVE-KD.
****

CamFreeDiff: Camera-free Image to Panorama Generation with Diffusion Model
Xiaoding Yuan, Shitao Tang, Kejie Li, Peng Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16408-16417
This paper introduces Camera-free Diffusion (CamFreeDiff) model for 360^\circ image outpainting from a single camera-free image and text description. This metho

d distinguishes itself from existing strategies, such as MVDiffusion, by elimina ting the requirement for predefined camera poses. CamFreeDiff seamlessly incorpo rates a mechanism for predicting homography within the multi-view diffusion fram ework. The key component of our approach is to formulate camera estimation by di rectly predicting the homography transformation from the input view to the prede fined canonical view. In contrast to the direct two-stage approach of image tran sformation and outpainting, CamFreeDiff utilizes predicted homography to establi sh point-level correspondences between the input view and the target panoramic v iew. This enables consistency through correspondence-aware attention, which is l earned in a fully differentiable manner. Qualitative and quantitative experiment al results demonstrate the strong robustness and performance of CamFreeDiff for 360^\circ image outpainting in the challenging context of camera-free inputs.
****

Improving Visual and Downstream Performance of Low-Light Enhancer with Vision Fo undation Models Collaboration

Yuxuan Gu, Haoxuan Wang, Pengyang Ling, Zhixiang Wei, Huaian Chen, Yi Jin, Enhon g Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (C VPR), 2025, pp. 16071-16080

In this paper, we observe that the collaboration of various foundation models ca n perceive semantic and degraded information within images, thereby guiding the low-light enhancement process. Specifically, we propose a self-supervised low-li ght enhancement framework based on the multiple foundation models collaboration (dubbed FoCo), aimed at improving both the visual quality of enhanced images and the performance in high-level applications. At the feature level, FoCo leverage s the rich features from various foundation models to enhance the model's semant ic perception during training, thereby reducing the gap between enhanced results and high-quality images from a high-level perspective. At the task level, we ex ploit the robustness-gap between strong foundation models and weak models, apply ing high-level task guidance to the low-light enhancement training process. Thro ugh the collaboration of multiple foundation models, the proposed framework show s better enhancement performance and adapts better to high-level tasks. Extensiv e experiments across various enhancement and application benchmarks demonstrate the qualitative and quantitative superiority of the proposed method over numerou s state-of-the-art techniques.
****

EchoWorld: Learning Motion-Aware World Models for Echocardiography Probe Guidanc e

Yang Yue, Yulin Wang, Haojun Jiang, Pan Liu, Shiji Song, Gao Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2599 3-26003

Echocardiography is crucial for cardiovascular disease detection but relies heav ily on experienced sonographers. Echocardiography probe guidance systems, which provide real-time movement instructions for acquiring standard plane images, off er a promising solution for AI-assisted or fully autonomous scanning. However, d eveloping effective machine learning models for this task remains challenging, a s they must grasp heart anatomy and the intricate interplay between probe motion and visual signals. To address this, we present EchoWorld, a motion-aware world modeling framework for probe guidance that encodes anatomical knowledge and mot ion-induced visual dynamics, while effectively leveraging past visual-motion seq uences to enhance guidance precision. EchoWorld employs a pre-training strategy inspired by world modeling principles, where the model predicts masked anatomica l regions and simulates the visual outcomes of probe adjustments. Built upon thi s pre-trained model, we introduce a motion-aware attention mechanism in the fine -tuning stage that effectively integrates historical visual-motion data, enablin g precise and adaptive probe guidance. Trained on more than one million ultrasou nd images from over 200 routine scans, EchoWorld effectively captures key echoca rdiographic knowledge, as validated by qualitative analysis. Moreover, our metho d significantly reduces guidance errors compared to existing visual backbones an d guidance frameworks, excelling in both single-frame and sequential evaluation protocols. Code is available at https://github.com/LeapLabTHU/EchoWorld.

****
Controllable Human Image Generation with Personalized Multi-Garments
Yisol Choi, Sangkyung Kwak, Sihyun Yu, Hyungwon Choi, Jinwoo Shin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28736-28747
We present BootControl, a novel framework based on text-to-image diffusion models for controllable human image generation with multiple reference garments.Here, the main bottleneck is data acquisition for training: collecting a large-scale dataset of high-quality reference garment images per human subject is quite challenging, i.e., ideally, one needs to manually gather every single garment photograph worn by each human.To address this, we propose a data generation pipeline to construct a large synthetic dataset, consisting of human and multiple-garment pairs, by introducing a model to extract any reference garment images from each human image.To ensure data quality, we also propose a filtering strategy to remove undesirable generated data based on measuring perceptual similarities between the garment presented in human image and extracted garment.Finally, by utilizing the constructed synthetic dataset, we train a diffusion model having two parallel denoising paths that use multiple garment images as conditions to generate human images while preserving their fine-grained details.We further show the wide-applicability of our framework by adapting it to different types of reference-based generation in the fashion domain, including virtual try-on, and controllable human image generation with other conditions, e.g., pose, face, etc.
****
FineLIP: Extending CLIP's Reach via Fine-Grained Alignment with Longer Text Inputs
Mothilal Asokan, Kebin Wu, Fatima Albreiki; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14495-14504
As a pioneering vision-language model, CLIP (Contrastive Language-Image Pre-training) has achieved significant success across various domains and a wide range of downstream vision-language tasks. However, the text encoders in popular CLIP models are limited to processing only 77 text tokens, which constrains their ability to effectively handle longer, detail-rich captions. Additionally, CLIP models often struggle to effectively capture detailed visual and textual information, which hampers their performance on tasks that require fine-grained analysis. To address these limitations, we present a novel approach, FineLIP, that extends the capabilities of CLIP. FineLIP enhances cross-modal text-image mapping by incorporating Fine-grained alignment with Longer text input within the CLIP-style framework. FineLIP first extends the positional embeddings to handle longer text, followed by the dynamic aggregation of local image and text tokens. The aggregated results are then used to enforce fine-grained token-to-token cross-modal alignment. We validate our model on datasets with long, detailed captions across two tasks: zero-shot cross-modal retrieval and text-to-image generation. Quantitative and qualitative experimental results demonstrate the effectiveness of FineLIP, outperforming existing state-of-the-art approaches. Furthermore, comprehensive ablation studies validate the benefits of key design elements within FineLIP. The code will be available at https://github.com/tiiuae/FineLIP.
****
Illumination Spectrum Estimation for Multispectral Images via Surface Reflectance Modeling and Spatial-Spectral Feature Generation
Hyejin Oh, Woo-Shik Kim, Sangyoon Lee, YungKyung Park, Je-Won Kang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2215-2225
Multispectral (MS) images contain richer spectral information than RGB images due to their increased number of channels and are widely used for various applications. However, achieving accurate estimation in MS images remains challenging, as previous studies have struggled with spectral diversity and the inherent entanglement between the illuminant and surface reflectance spectra. To tackle these challenges, in this paper, we propose a novel Illumination spectrum estimation technique for MS images via Surface reflectance modeling and Spatial-spectral feature generation (ISS). The proposed technique employs a learnable spectral unmix

ing (SU) block to enhance surface reflectance modeling, which was unattempted in the illumination spectrum estimation, and a feature mixing block to fuse spectral and spatial features of MS images with cross-attention. The features are refined iteratively and processed through a decoder to produce an illumination spectrum estimator. Experimental results demonstrate that the proposed technique achieves state-of-the-art performance in illumination spectrum estimation in various MS image datasets. The code is available at https://github.com/heyjinnii/ISS-MSI.git.

****

UHD-processer: Unified UHD Image Restoration with Progressive Frequency Learning and Degradation-aware Prompts

Yidi Liu, Dong Li, Xueyang Fu, Xin Lu, Jie Huang, Zheng-Jun Zha; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23121-23130

We introduce UHD-Processor, a unified and robust framework for all-in-one image restoration, which is particularly resource-efficient for Ultra-High-Definition (UHD) images. To address the limitations of traditional all-in-one methods that rely on complex restoration backbones, our strategy employs a frequency domain decoupling progressive learning technique, motivated by curriculum learning, to incrementally learn restoration mappings from low to high frequencies. This approach incorporates specialized sub-network modules to effectively tackle different frequency bands in a divide-and-conquer manner, significantly enhancing the learning capability of simpler networks. Moreover, to accommodate the high-resolution characteristics of UHD images, we developed a variational autoencoder (VAE)-based framework that reduces computational complexity by modeling a concise latent space. It integrates task-specific degradation awareness in the encoder and frequency selection in the decoder, enhancing task comprehension and generalization. Our unified model is able to handle various degradations such as denoising, deblurring, dehazing, low-lighting, etc. Experimental evaluations extensively showcase the effectiveness of our dual-strategy approach, significantly improving UHD image restoration and achieving cutting-edge performance across diverse conditions. The code will be available at https://github.com/lyd-2022/UHD-processer.

****

Divot: Diffusion Powers Video Tokenizer for Comprehension and Generation

Yuying Ge, Yizhuo Li, Yixiao Ge, Ying Shan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13606-13617

In recent years, there has been a significant surge of interest in unifying image comprehension and generation within Large Language Models (LLMs). This growing interest has prompted us to explore extending this unification to videos. The core challenge lies in developing a versatile video tokenizer that captures both the spatial characteristics and temporal dynamics of videos to obtain representations for LLMs, and the representations can be further decoded into realistic video clips to enable video generation. In this work, we introduce Divot, a Diffusion-Powered Video Tokenizer, which leverages the diffusion process for self-supervised video representation learning. We posit that if a video diffusion model can effectively de-noise video clips by taking the features of a video tokenizer as the condition, then the tokenizer has successfully captured robust spatial and temporal information. Additionally, the video diffusion model inherently functions as a de-tokenizer, decoding videos from their representations. Building upon the Divot tokenizer, we present Divot-LLM through video-to-text autoregression and text-to-video generation by modeling the distributions of continuous-valued Divot features with a Gaussian Mixture Model. Experimental results demonstrate that our diffusion-based video tokenizer, when integrated with a pre-trained LLM, achieves competitive performance across various video comprehension and generation benchmarks. The instruction tuned Divot-LLM also excels in video storytelling, generating interleaved narratives and corresponding videos. Models and codes are available at https://github.com/TencentARC/Divot.

****

Towards Zero-Shot Anomaly Detection and Reasoning with Multimodal Large Language Models

Jiacong Xu, Shao-Yuan Lo, Bardia Safaei, Vishal M. Patel, Isht Dwivedi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20370-20382

Zero-Shot Anomaly Detection (ZSAD) is an emerging AD paradigm. Unlike the traditional unsupervised AD setting that requires a large number of normal samples to train a model, ZSAD is more practical for handling data-restricted real-world scenarios. Recently, Multimodal Large Language Models (MLLMs) have shown revolutionary reasoning capabilities in various vision tasks. However, the reasoning of image abnormalities remains underexplored due to the lack of corresponding datasets and benchmarks. To facilitate research in AD & reasoning, we establish the first visual instruction tuning dataset, Anomaly-Instruct-125k, and the evaluation benchmark, VisA-D&R. Through investigation with our benchmark, we reveal that current MLLMs like GPT-4o cannot accurately detect and describe fine-grained anomalous details in images. To address this, we propose Anomaly-OneVision (Anomaly-OV), the first specialist visual assistant for ZSAD and reasoning. Inspired by human behavior in visual inspection, Anomaly-OV leverages a Look-Twice Feature Matching (LTFM) mechanism to adaptively select and emphasize abnormal visual tokens. Extensive experiments demonstrate that Anomaly-OV achieves significant improvements over advanced generalist models in both detection and reasoning. Extensions to medical and 3D AD are provided for future study. The link to our project page: https://xujiacong.github.io/Anomaly-OV/

****

Towards Explainable and Unprecedented Accuracy in Matching Challenging Finger Crease Patterns

Zhenyu Zhou, Chengdong Dong, Ajay Kumar; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6212-6221

The primary obstacle in realizing the full potential of finger crease biometrics is the accurate identification of deformed knuckle patterns, often resulting from completely contactless imaging. Current methods struggle significantly with this task, yet accurate matching is crucial for applications ranging from forensic investigations, such as child abuse cases, to surveillance and mobile security. To address this challenge, our study introduces the largest publicly available dataset of deformed knuckle patterns, comprising 805,768 images from 351 subjects. We also propose a novel framework to accurately match knuckle patterns, even under severe pose deformations, by recovering interpretable knuckle crease keypoint feature templates. These templates can dynamically uncover graph structure and feature similarity among the matched correspondences. Our experiments, using the most challenging protocols, illustrate significantly outperforming results for matching such knuckle images. For the first time, we present and evaluate a theoretical model to estimate the uniqueness of 2D finger knuckle patterns, providing a more interpretable and accurate measure of distinctiveness, which is invaluable for forensic examiners in prosecuting suspects.

****

Neural Hierarchical Decomposition for Single Image Plant Modeling

Zhihao Liu, Zhanglin Cheng, Naoto Yokoya; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 733-742

Obtaining high-quality, practically usable 3D models of biological plants remains a significant challenge in computer vision and graphics. In this paper, we present a novel method for generating realistic 3D plant models from single-view photographs. Our approach employs a neural decomposition technique to learn a lightweight hierarchical box representation from the image, effectively capturing the structures and botanical features of plants. Then, this representation can be subsequently refined through a shape-guided parametric modeling module to produce complete 3D plant models. By combining hierarchical learning and parametric modeling, our method generates structured 3D plant assets with fine geometric details. Notably, through learning the decomposition in different levels of detail, our method can adapt to two distinct plant categories: outdoor trees and houseplants, each with unique appearance features. Within the scope of plant modeling, our method is the first comprehensive solution capable of reconstructing both plant categories from single-view images.

****
GBC-Splat: Generalizable Gaussian-Based Clothed Human Digitalization under Sparse RGB Cameras
Hanzhang Tu, Zhanfeng Liao, Boyao Zhou, Shunyuan Zheng, Xilong Zhou, Liuxin Zhang, QianYing Wang, Yebin Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26377-26387
We present an efficient approach for generalizable clothed human digitalization, termed GBC-Splat. Unlike previous methods that necessitate per-subject optimizations or discount watertight geometry, the proposed method is dedicated to reconstructing complete human shapes and Gaussian Splatting via sparse view RGB inputs in a feed-forward manner. We first extract a fine-grained mesh using a combination of implicit occupancy field regression and explicit disparity estimation between views. Gaussian primitives anchored on the mesh allow 6-DoF photorealistic view synthesis. The reconstructed high-quality geometry allows us to easily anchor Gaussian primitives to mesh surface according to surface normal and texture, which allows 6-DoF photorealistic novel view synthesis. In addition, we introduce a simple yet effective algorithm to subdivide Gaussian primitives in high-frequency areas to further enhance the visual quality. Without the assistance of human parametric models, our method can tackle loose garments, such as dresses and costumes. Our method outperforms state-of-the-art methods in terms of novel view synthesis while keeping high efficiency, enabling the potential of deployment in real-time applications.
****
AC3D: Analyzing and Improving 3D Camera Control in Video Diffusion Transformers
Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B. Lindell, Sergey Tulyakov; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22875-22889
Numerous works have recently integrated 3D camera control into foundational text-to-video models, but the resulting camera control is often imprecise, and video generation quality suffers. In this work, we analyze camera motion from a first principles perspective, uncovering insights that enable precise 3D camera manipulation without compromising synthesis quality. First, we determine that motion induced by camera movements in videos is low-frequency in nature. This motivates us to adjust train and test pose conditioning schedules, accelerating training convergence while improving visual and motion quality. Then, by probing the representations of an unconditional video diffusion transformer, we observe that they implicitly perform camera pose estimation under the hood, and only a sub-portion of their layers contain the camera information. This suggested us to limit the injection of camera conditioning to a subset of the architecture to prevent interference with other video features, leading to a 4x reduction of training parameters, improved training speed, and 10% higher visual quality. Finally, we complement the typical dataset for camera control learning with a curated dataset of 20K diverse, dynamic videos with stationary cameras. This helps the model distinguish between camera and scene motion and improves the dynamics of generated pose-conditioned videos. We compound these findings to design the Advanced 3D Camera Control (AC3D) architecture, the new state-of-the-art model for generative video modeling with camera control.
****
A Unified Model for Compressed Sensing MRI Across Undersampling Patterns
Armeet Singh Jatyani, Jiayun Wang, Aditi Chandrashekar, Zihui Wu, Miguel Liu-Schiaffini, Bahareh Tolooshams, Anima Anandkumar; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26004-26013
Compressed Sensing MRI reconstructs images of the body's internal anatomy from undersampled measurements, thereby reducing the scan time - the time subjects need to remain still. Recently, deep learning has shown great potential for reconstructing high-fidelity images from highly undersampled measurements. However, one needs to train multiple models for different undersampling patterns and desired output image resolutions, since most networks operate on a fixed discretization. Such approaches are highly impractical in clinical settings, where undersampl

ing patterns and image resolutions are frequently changed to accommodate differe nt real-time imaging and diagnostic requirements. We propose a unified MRI recon struction model robust to various measurement undersampling patterns and image r esolutions. Our approach uses neural operators- a discretization-agnostic archi tecture applied in both image and measurement spaces--to capture local and globa l features. Empirically, our model improves SSIM by 11% and PSNR by 4dB over a s tate-of-the-art CNN (End-to-End VarNet), with inference 600x faster than diffusi on methods. The resolution-agnostic design also enables zero-shot super-resoluti on and extended field-of-view reconstruction, offering a versatile and efficient solution for clinical MR imaging. Our unified model offers a versatile solution for MRI, adapting seamlessly to various measurement undersampling and imaging r esolutions, making it highly effective for flexible and reliable clinical imagin g. Our code is available at https://armeet.ca/nomri.
****

Video-Guided Foley Sound Generation with Multimodal Controls
Ziyang Chen, Prem Seetharaman, Bryan Russell, Oriol Nieto, David Bourgin, Andrew Owens, Justin Salamon; Proceedings of the Computer Vision and Pattern Recogniti on Conference (CVPR), 2025, pp. 18770-18781
Generating sound effects for videos often requires creating artistic sound effec ts that diverge significantly from real-life sources and flexible control in the sound design. To address this problem, we introduce *MultiFoley*, a model desig ned for video-guided sound generation that supports multimodal conditioning thro ugh text, audio, and video. Given a silent video and a text prompt, MultiFoley a llows users to create clean sounds (e.g., skateboard wheels spinning without win d noise) or more whimsical sounds (e.g., making a lion's roar sound like a cat's meow).MultiFoley also allows users to choose reference audio from sound effects (SFX) libraries or partial videos for conditioning. A key novelty of our model lies in its joint training on both internet video datasets with low-quality audi o and professional SFX recordings, enabling high-quality, full-bandwidth (48kHz) audio generation.Through automated evaluations and human studies, we demonstrat e that *MultiFoley* successfully generates synchronized high-quality sounds acro ss varied conditional inputs and outperforms existing methods.
****

Dual-Agent Optimization framework for Cross-Domain Few-Shot Segmentation
Zhaoyang Li, Yuan Wang, Wangkai Li, Tianzhu Zhang, Xiang Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9849-9859
Cross-Domain Few-Shot Segmentation (CD-FSS) extends the generalization ability o f Few-Shot Segmentation (FSS) beyond a single domain, enabling more practical ap plications. However, directly employing conventional FSS methods suffers from se vere performance degradation in cross-domain settings, primarily due to feature sensitivity and support-to-query matching process sensitivity across domains. Ex isting methods for CD-FSS either focus on domain adaptation of features or delve into designing matching strategies for enhanced cross-domain robustness. Noneth eless, they overlook the fact that these two issues are interdependent and shoul d be addressed jointly. In this work, we tackle these two issues within a unifie d framework by optimizing features in the frequency domain and enhancing the mat ching process in the spatial domain, working jointly to handle the deviations in troduced by the domain gap. To this end, we propose a coherent Dual-Agent Optimi zation (DATO) framework, including a consistent mutual aggregation (CMA) and a c orrelation rectification strategy (CRS). In the consistent mutual aggregation mo dule, we employ a set of agents to learn domain-invariant features across domain s, and then use these features to enhance the original representations for featu re adaptation. In the correlation rectification strategy, the agent-aggregated d omain-invariant features serve as a bridge, transforming the support-to-query ma tching process into a referable feature space and reducing its domain sensitivit y. Extensive experiments demonstrate the efficacy of our approach.
****

SACB-Net: Spatial-awareness Convolutions for Medical Image Registration
Xinxing Cheng, Tianyang Zhang, Wenqi Lu, Qingjie Meng, Alejandro F. Frangi, Jinm ing Duan; Proceedings of the Computer Vision and Pattern Recognition Conference

(CVPR), 2025, pp. 5227-5237
Deep learning-based image registration methods have shown state-of-the-art perfo rmance and rapid inference speeds. Despite these advances, many existing approac hes fall short in capturing spatially varying information in non-local regions o f feature maps due to the reliance on spatially-shared convolution kernels. This limitation leads to suboptimal estimation of deformation fields. In this paper, we propose a 3D Spatial-Awareness Convolution Block (SACB) to enhance the spati al information within feature representations. Our SACB estimates the spatial cl usters within feature maps by leveraging feature similarity and subsequently par ameterizes the adaptive convolution kernels across diverse regions. This adaptiv e mechanism generates the convolution kernels (weights and biases) tailored to s patial variations, thereby enabling the network to effectively capture spatially varying information. Building on SACB, we introduce a pyramid flow estimator (n amed SACB-Net) that integrates SACBs to facilitate multi-scale flow composition, particularly addressing large deformations. Experimental results on the brain I XI and LPBA datasets as well as Abdomen CT datasets demonstrate the effectivenes s of SACB and the superiority of SACB-Net over the state-of-the-art learning-bas ed registration methods. The code is available at https://github.com/x-xc/SACB_ Net.
****

Text Embedding is Not All You Need: Attention Control for Text-to-Image Semantic Alignment with Text Self-Attention Maps
Jeeyung Kim, Erfan Esmaeili, Qiang Qiu; Proceedings of the Computer Vision and P attern Recognition Conference (CVPR), 2025, pp. 8031-8040
In text-to-image diffusion models, the cross-attention map of each text token in dicates the specific image regions attended. Comparing these maps of syntactical ly related tokens provides insights into how well the generated image reflects t he text prompt. For example, in the prompt, "a black car and a white clock", the cross-attention maps for "black" and "car" should focus on overlapping regions to depict a black car, while "car" and "clock" should not. Incorrect overlapping in the maps generally produces generation flaws such as missing objects and inc orrect attribute binding. Our study makes the key observations investigating thi s issue in the existing text-to-image models: (1) the similarity in text embeddi ngs between different tokens---used as conditioning inputs---can cause their cro ss-attention maps to focus on the same image regions; and (2) text embeddings of ten fail to faithfully capture syntactic relations already within text attention maps. As a result, such syntactic relationships can be overlooked in cross-atte ntion module, leading to inaccurate image generation. To address this, we propos e a method that directly transfers syntactic relations from the text attention m aps to the cross-attention module via a test-time optimization. Our approach lev erages this inherent yet unexploited information within text attention maps to e nhance image-text semantic alignment across diverse prompts, without relying on external guidance. Our project page and code are available at: https://t-sam-dif fusion.github.io
****

DCEvo: Discriminative Cross-Dimensional Evolutionary Learning for Infrared and V isible Image Fusion
Jinyuan Liu, Bowei Zhang, Qingyun Mei, Xingyuan Li, Yang Zou, Zhiying Jiang, Lon g Ma, Risheng Liu, Xin Fan; Proceedings of the Computer Vision and Pattern Recog nition Conference (CVPR), 2025, pp. 2226-2235
Infrared and visible image fusion integrates information from distinct spectral bands to enhance image quality by leveraging the strengths and mitigating the li mitations of each modality. Existing approaches typically treat image fusion and subsequent high-level tasks as separate processes, resulting in fused images th at offer only marginal gains in task performance and fail to provide constructiv e feedback for optimizing the fusion process. To overcome these limitations, we propose a Discriminative Cross-Dimension Evolutionary Learning Framework, termed DCEvo, which simultaneously enhances visual quality and perception accuracy. Le veraging the robust search capabilities of Evolutionary Learning, our approach f ormulates the optimization of dual tasks as a multi-objective problem by employi

ng an Evolutionary Algorithm (EA) to dynamically balance loss function parameters. Inspired by visual neuroscience, we integrate a Discriminative Enhancer (DE) within both the encoder and decoder, enabling the effective learning of complementary features from different modalities. Additionally, our Cross-Dimensional Embedding (CDE) block facilitates mutual enhancement between high-dimensional task features and low-dimensional fusion features, ensuring a cohesive and efficient feature integration process. Experimental results on three benchmarks demonstrate that our method significantly outperforms state-of-the-art approaches, achieving an average improvement of 9.32% in visual quality while also enhancing subsequent high-level tasks. The code is available at https://github.com/Beate-Suy-Zhang/DCEvo.

****

TSD-SR: One-Step Diffusion with Target Score Distillation for Real-World Image Super-Resolution

Linwei Dong, Qingnan Fan, Yihong Guo, Zhonghao Wang, Qi Zhang, Jinwei Chen, Yawei Luo, Changqing Zou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23174-23184

Pre-trained text-to-image diffusion models are increasingly applied to real-world image super-resolution (Real-ISR) task. Given the iterative refinement nature of diffusion models, most existing approaches are computationally expensive. While methods such as SinSR and OSEDiff have emerged to condense inference steps via distillation, their performance in image restoration or details recovery is not satisfied. To address this, we propose TSD-SR, a novel distillation framework specifically designed for real-world image super-resolution, aiming to construct an efficient and effective one-step model. We first introduce the Target Score Distillation, which leverages the priors of diffusion models and real image references to achieve more realistic image restoration. Secondly, we propose a Distribution-Aware Sampling Module to make detail-oriented gradients more readily accessible, addressing the challenge of recovering fine details. Extensive experiments demonstrate that our TSD-SR has superior restoration results (most of the metrics perform the best) and the fastest inference speed (e.g. 40 times faster than SeeSR) compared to the past Real-ISR approaches based on pre-trained diffusion priors.

****

AIpparel: A Multimodal Foundation Model for Digital Garments

Kiyohiro Nakayama, Jan Ackermann, Timur Levent Kesdogan, Yang Zheng, Maria Korosteleva, Olga Sorkine-Hornung, Leonidas J. Guibas, Guandao Yang, Gordon Wetzstein; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8138-8149

Apparel is essential to human life, offering protection, mirroring cultural identities, and showcasing personal style. Yet, the creation of garments remains a time-consuming process, largely due to the manual work involved in designing them. To simplify this process, we introduce AIpparel, a multimodal foundation model for generating and editing sewing patterns. Our model fine-tunes state-of-the-art large multimodal models (LMMs) on a custom-curated large-scale dataset of over 120,000 unique garments, each with multimodal annotations including text, images, and sewing patterns. Additionally, we propose a novel tokenization scheme that concisely encodes these complex sewing patterns so that LLMs can learn to predict them efficiently. AIpparel achieves state-of-the-art performance in single-modal tasks, including text-to-garment and image-to-garment prediction, and enables novel multimodal garment generation applications such as interactive garment editing. The project website is at https://georgenakayama.github.io/AIpparel/.

****

Fast3R: Towards 3D Reconstruction of 1000+ Images in One Forward Pass

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, Matt Feiszli; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21924-21935

Multi-view 3D reconstruction remains a core challenge in computer vision, particularly in applications requiring accurate and scalable representations across diverse perspectives. Current leading methods such as DUSt3R employ a fundamentall

y pairwise approach, processing images in pairs and necessitating costly global alignment procedures to reconstruct from multiple views. In this work, we propose Fast 3D Reconstruction (Fast3R), a novel multi-view generalization to DUSt3R that achieves efficient and scalable 3D reconstruction by processing many views in parallel. Fast3R's Transformer-based architecture forwards N images in a single forward pass, bypassing the need for iterative alignment. Through extensive experiments on camera pose estimation and 3D reconstruction, Fast3R demonstrates state-of-the-art performance, with significant improvements in inference speed and reduced error accumulation. These results establish Fast3R as a robust alternative for multi-view applications, offering enhanced scalability without compromising reconstruction accuracy. Project website: https://fast3r-3d.github.io
****

StyleStudio: Text-Driven Style Transfer with Selective Control of Style Elements
Mingkun Lei, Xue Song, Beier Zhu, Hao Wang, Chi Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23443-23452
Text-driven style transfer aims to merge the style of a reference image with content described by a text prompt. Recent advancements in text-to-image models have improved the nuance of style transformations, yet significant challenges remain, particularly with overfitting to reference styles, limiting stylistic control, and misaligning with textual content.In this paper, we propose three complementary strategies to address these issues. First, we introduce a cross-modal Adaptive Instance Normalization (AdaIN) mechanism for better integration of style and text features, enhancing alignment. Second, we develop a Style-based Classifier-Free Guidance (SCFG) approach that enables selective control over stylistic elements, reducing irrelevant influences. Finally, we incorporate a teacher model during early generation stages to stabilize spatial layouts and mitigate artifacts. Our extensive evaluations demonstrate significant improvements in style transfer quality and alignment with textual prompts. Furthermore, our approach can be integrated into existing style transfer frameworks without fine-tuning.
****

CTRL-O: Language-Controllable Object-Centric Visual Representation Learning
Aniket Didolkar, Andrii Zadaianchuk, Rabiul Awal, Maximilian Seitzer, Efstratios Gavves, Aishwarya Agrawal; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29523-29533
Object-centric representation learning aims to decompose visual scenes into fixed-size vectors called "slots" or "object files", where each slot captures a distinct object. Current state-of-the-art object-centric models have shown remarkable success in object discovery in diverse domains including complex real-world scenes. However, these models suffer from a key limitation: they lack controllability. Specifically, current object-centric models learn representations based on their preconceived understanding of objects and parts, without allowing user input to guide which objects are represented. Introducing controllability into object-centric models could unlock a range of useful capabilities, such as the ability to extract instance-specific representations from a scene. In this work, we propose a novel approach for user-directed control over slot representations by conditioning slots on language descriptions. The proposed ConTRoLlable Object-centric representation learning approach, which we term CTRL-O, achieves targeted object-language binding in complex real-world scenes without requiring mask supervision. Next, we apply these controllable slot representations on two downstream vision language tasks: text-to-image generation and visual question answering. We find that the proposed approach enables instance-specific text-to-image generation and also achieves strong performance on visual question answering.
****

PO3AD: Predicting Point Offsets toward Better 3D Point Cloud Anomaly Detection
Jianan Ye, Weiguang Zhao, Xi Yang, Guangliang Cheng, Kaizhu Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1353-1362
Point cloud anomaly detection under the anomaly-free setting poses significant challenges as it requires accurately capturing the features of 3D normal data to identify deviations indicative of anomalies. Current efforts focus on devising r

econstruction tasks, such as acquiring normal data representations by restoring normal samples from altered, pseudo-anomalous counterparts. Our findings reveal that distributing attention equally across normal and pseudo-anomalous data tends to dilute the model's focus on anomalous deviations. The challenge is further compounded by the inherently disordered and sparse nature of 3D point cloud data. In response to those predicaments, we introduce an innovative approach that emphasizes learning point offsets, targeting more informative pseudo-abnormal points, thus fostering more effective distillation of normal data representations. We also have crafted an augmentation technique that is steered by normal vectors, facilitating the creation of credible pseudo anomalies that enhance the efficiency of the training process. Our comprehensive experimental evaluation on the Anomaly-ShapeNet and Real3D-AD datasets evidences that our proposed method outperforms existing state-of-the-art approaches, achieving an average enhancement of 9.0% and 1.4% in the AUC-ROC detection metric across these datasets, respectively. Code is available at https://github.com/yjnanan/PO3AD.

****

Text Augmented Correlation Transformer For Few-shot Classification & Segmentation

Srinivasa Rao Nandam, Sara Atito, Zhenhua Feng, Josef Kittler, Muhammad Awais; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25357-25366

Foundation models like CLIP and ALIGN have transformed few-shot and zero-shot vision applications by fusing visual and textual data, yet the integrative few-shot classification and segmentation (FS-CS) task primarily leverages visual cues, overlooking the potential of textual support. In FS-CS scenarios, ambiguous object boundaries and overlapping classes often hinder model performance, as limited visual data struggles to fully capture high-level semantics. To bridge this gap, we present a novel multi-modal FS-CS framework that integrates textual cues into support data, facilitating enhanced semantic disambiguation and fine-grained segmentation. Our approach first investigates the unique contributions of exclusive text-based support, using only class labels to achieve FS-CS. This strategy alone achieves performance competitive with vision-only methods on FS-CS tasks, underscoring the power of textual cues in few-shot learning. Building on this, we introduce a dual-modal prediction mechanism that synthesizes insights from both textual and visual support sets, yielding robust multi-modal predictions. This integration significantly elevates FS-CS performance, with classification and segmentation improvements of +3.7/6.6% (1-way 1-shot) and +8.0/6.5% (2-way 1-shot) on COCO-20^i, and +2.2/3.8% (1-way 1-shot) and +4.3/4.0% (2-way 1-shot) on Pascal-5^i. Additionally, in weakly supervised FS-CS settings, our method surpasses visual-only benchmarks using textual support exclusively, further enhanced by our dual-modal predictions. By rethinking the role of text in FS-CS, our work establishes new benchmarks for multi-modal few-shot learning and demonstrates the efficacy of textual cues for improving model generalization and segmentation accuracy.

****

F^3OCUS - Federated Finetuning of Vision-Language Foundation Models with Optimal Client Layer Updating Strategy via Multi-objective Meta-Heuristics

Pramit Saha, Felix Wagner, Divyanshu Mishra, Can Peng, Anshul Thakur, David A. Clifton, Konstantinos Kamnitsas, J. Alison Noble; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20006-20017

Effective training of large Vision-Language Models (VLMs) on resource-constrained client devices in Federated Learning (FL) requires the usage of parameter-efficient fine-tuning (PEFT) strategies. To this end, we demonstrate the impact of two factors, viz., client-specific layer importance score that selects the most important VLM layers for fine-tuning and inter-client layer diversity score that encourages diverse layer selection across clients for optimal VLM layer selection. We first theoretically motivate and leverage the principal eigenvalue magnitude of layerwise Neural Tangent Kernels and show its effectiveness as client-specific layer importance score. Next, we propose a novel layer updating strategy dubbed F^3OCUS that jointly optimizes the layer importance and diversity factors b

y employing a data-free, multi-objective, meta-heuristic optimization on the ser ver. We explore 5 different meta-heuristic algorithms and compare their effectiv eness for selecting model layers and adapter layers towards PEFT-FL. Furthermore , we release a new MedVQA-FL dataset involving overall 707,962 VQA triplets and 9 modality-specific clients and utilize it to train and evaluate our method. Ove rall, we conduct more than 10,000 client-level experiments on 6 Vision-Language FL task settings involving 58 medical image datasets and 4 different VLM archite ctures of varying sizes to demonstrate the effectiveness of the proposed method. Project Page: https://pramitsaha.github.io/FOCUS/

****

ICT: Image-Object Cross-Level Trusted Intervention for Mitigating Object Halluci nation in Large Vision-Language Models

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, Xu ming Hu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4209-4221

Despite the recent breakthroughs achieved by Large Vision Language Models (LVLMs) in understanding and responding to complex visual-textual contexts, their inhe rent hallucination tendencies limit their practical application in real-world sc enarios that demand high levels of precision. Existing methods typically either fine-tune the LVLMs using additional data, which incurs extra costs in manual an notation and computational resources or perform comparisons at the decoding stag e, which may eliminate useful language priors for reasoning while introducing in ference time overhead. Therefore, we propose ICT, a lightweight, training-free m ethod that calculates an intervention direction to shift the model's focus towar ds different levels of visual information, enhancing its attention to high-level and fine-grained visual details. During the forward pass stage, the interventi on is applied to the attention heads that encode the overall image information a nd the fine-grained object details, effectively mitigating the phenomenon of ove rly language priors, and thereby alleviating hallucinations. Extensive experimen ts demonstrate that ICT achieves strong performance with a small amount of data and generalizes well across different datasets and models. Our code will be publ ic.

****

PreciseCam: Precise Camera Control for Text-to-Image Generation

Edurne Bernal-Berdun, Ana Serrano, Belen Masia, Matheus Gadelha, Yannick Hold-Ge offroy, Xin Sun, Diego Gutierrez; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2724-2733

Images as an artistic medium often rely on specific camera angles and lens disto rtions to convey ideas or emotions; however, such precise control is missing in current text-to-image models. We propose an efficient and general solution that allows precise control over the camera when generating both photographic and art istic images. Unlike prior methods that rely on predefined shots, we rely solely on four simple extrinsic and intrinsic camera parameters, removing the need for pre-existing geometry, reference 3D objects, and multi-view data.We also presen t a novel dataset with more than 57,000 images, along with their text prompts an d ground-truth camera parameters. Our evaluation shows precise camera control in text-to-image generation, surpassing traditional prompt engineering approaches.

****

3D Occupancy Prediction with Low-Resolution Queries via Prototype-aware View Tra nsformation

Gyeongrok Oh, Sungjune Kim, Heeju Ko, Hyung-gun Chi, Jinkyu Kim, Dongwook Lee, D aehyun Ji, Sungjoon Choi, Sujin Jang, Sangpil Kim; Proceedings of the Computer V ision and Pattern Recognition Conference (CVPR), 2025, pp. 17134-17144

The resolution of voxel queries significantly influences the quality of view tra nsformation in camera-based 3D occupancy prediction. However, computational cons traints and the practical necessity for real-time deployment require smaller que ry resolutions, which inevitably leads to an information loss. Therefore, it is essential to encode and preserve rich visual details within limited query sizes while ensuring a comprehensive representation of 3D occupancy. To this end, we i ntroduce ProtoOcc, a novel occupancy network that leverages prototypes of cluste

red image segments in view transformation to enhance low-resolution context. In particular, the mapping of 2D prototypes onto 3D voxel queries encodes high-level visual geometries and complements the loss of spatial information from reduced query resolutions. Additionally, we design a multi-perspective decoding strategy to efficiently disentangle the densely compressed visual cues into a high-dimensional 3D occupancy scene. Experimental results on both Occ3D and SemanticKITTI benchmarks demonstrate the effectiveness of the proposed method, showing clear improvements over the baselines. More importantly, ProtoOcc achieves competitive performance against the baselines even with 75% reduced voxel resolution.
****

Unified Dense Prediction of Video Diffusion

Lehan Yang, Lu Qi, Xiangtai Li, Sheng Li, Varun Jampani, Ming-Hsuan Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28963-28973

We present a unified network for simultaneously generating videos and their corresponding entity segmentation and depth maps from text prompts. We utilize color map to represent entity masks and depth maps, tightly integrating dense prediction with RGB video generation. Introducing dense prediction information improves video generation's consistency and motion smoothness without increasing computational costs. Incorporating learnable task embeddings brings multiple dense prediction tasks into a single model, enhancing flexibility and further boosting performance. We further propose a large-scale dense prediction video dataset Panda-Dense, addressing the issue that existing datasets do not concurrently contain captions, videos, segmentation, or depth maps. Comprehensive experiments demonstrate the high efficiency of our method, surpassing the state-of-the-art in terms of video quality, consistency, and motion smoothness.
****

Can Large Vision-Language Models Correct Semantic Grounding Errors By Themselves?

Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, David Acuna; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14667-14678

Improving semantic grounding in Vision-Language Models (VLMs) often involves collecting domain-specific training data, refining the network architectures, or modifying the training recipes. In this work, we venture into an orthogonal direction and explore self-correction in VLMs focusing on semantic grounding. We find that VLMs can correct their own semantic grounding mistakes when properly prompted and framed for the task, without any fine-tuning or even access to oracle feedback. We also introduce a self-correction framework in an iterative setting which consistently improves performance across all models investigated. Overall, we show that iterative self-correction consistently improves VLM performance in semantic grounding by up to 8.4 accuracy points across all models investigated, without requiring fine-tuning, additional architectural changes, or external data. Our exploration of self-correction also reveals that, even after several rounds of feedback, strong models like GPT-4V and GPT-4o retain limited capability in leveraging oracle feedback, suggesting promising directions for further research.
****

SET: Spectral Enhancement for Tiny Object Detection

Huixin Sun, Runqi Wang, Yanjing Li, Linlin Yang, Shaohui Lin, Xianbin Cao, Baochang Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4713-4723

Deep learning has significantly advanced the object detection field. However, tiny object detection (TOD) remains a challenging problem. We provide a new analysis method to examine the TOD challenge through occlusion-based attribution analysis in the frequency domain. We observe that tiny objects become less distinct after feature encoding and can benefit from the removal of high-frequency information. In this paper, we propose a novel approach named Spectral Enhancement for Tiny object detection (SET), which amplifies the frequency signatures of tiny objects in a heterogeneous architecture. SET includes two modules. The Hierarchical Background Smoothing (HBS) module suppresses high-frequency noise in the backg

round through adaptive smoothing operations. The Adversarial Perturbation Injection (API) module leverages adversarial perturbations to increase feature saliency in critical regions and prompt the refinement of object features during training. Extensive experiments on four datasets demonstrate the effectiveness of our method. Especially, SET boosts the prior art RFLA by 3.2% AP on the AI-TOD dataset.

****

g3D-LF: Generalizable 3D-Language Feature Fields for Embodied Tasks

Zihan Wang, Gim Hee Lee; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14191-14202

We introduce Generalizable 3D-Language Feature Fields (g3D-LF), a 3D representation model pre-trained on large-scale 3D-language dataset for embodied tasks. Our g3D-LF processes posed RGB-D images from agents to encode feature fields for: 1) Novel view representation predictions from any position in the 3D scene; 2) Generations of BEV maps centered on the agent; 3) Querying targets using multi-granularity language within the above-mentioned representations. Our representation can be generalized to unseen environments, enabling real-time construction and dynamic updates. By volume rendering latent features along sampled rays and integrating semantic and spatial relationships through multiscale encoders, our g3D-LF produces representations at different scales and perspectives, aligned with multi-granularity language, via multi-level contrastive learning. Furthermore, we prepare a large-scale 3D-language dataset to align the representations of the feature fields with language. Extensive experiments on Vision-and-Language Navigation under both Panorama and Monocular settings, Zero-shot Object Navigation, and Situated Question Answering tasks highlight the significant advantages and effectiveness of our g3D-LF for embodied tasks. Our source code and dataset will be made open-source upon paper acceptance.

****

Towards Million-Scale Adversarial Robustness Evaluation With Stronger Individual Attacks

Yong Xie, Weijie Zheng, Hanxun Huang, Guangnan Ye, Xingjun Ma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30702-30711

As deep learning models are increasingly deployed in safety-critical applications, evaluating their vulnerabilities to adversarial perturbations is essential for ensuring their reliability and trustworthiness. Over the past decade, a large number of white-box adversarial robustness methods (i.e., attacks) have been proposed, ranging from single-step to multi-step methods and from individual to ensemble methods. Despite these advances, challenges remain in conducting meaningful and comprehensive robustness evaluations, particularly when it comes to large-scale testing and ensuring evaluations reflect real-world adversarial risks.In this work, we focus on image classification models and propose a novel individual attack method, Probability Margin Attack (PMA), which defines the adversarial margin in the probability space rather than the logits space. We analyze the relationship between PMA and existing cross-entropy or logits-margin-based attacks, showing that PMA outperforms the current state-of-the-art individual methods.Building on PMA, we propose two types of ensemble attacks that balance effectiveness and efficiency. Furthermore, we create a million-scale dataset, CC1M, derived from the existing CC3M dataset, and use it to conduct the first million-scale white-box adversarial robustness evaluation of adversarially-trained ImageNet models. Our findings provide valuable insights into the robustness gaps between individual versus ensemble attacks and small-scale versus million-scale evaluations.

****

Temporal Action Detection Model Compression by Progressive Block Drop

Xiaoyong Chen, Yong Guo, Jiaming Liang, Sitong Zhuang, Runhao Zeng, Xiping Hu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29225-29236

Temporal action detection (TAD) aims to identify and localize action instances in untrimmed videos, which is essential for various video understanding tasks. However, recent improvements in model performance, driven by larger feature extrac

tors and datasets, have led to increased computational demands. This presents a challenge for applications like autonomous driving and robotics, which rely on limited computational resources. While existing channel pruning methods can compress these models, reducing the number of channels often hinders the parallelization efficiency of GPU, due to the inefficient multiplication between small matrices. Instead of pruning channels, we propose a **Progressive Block Drop** method that reduces model depth while retaining layer width. In this way, we still use large matrices for computation but reduce the number of multiplications. Our approach iteratively removes redundant blocks in two steps: first, we drop blocks with minimal impact on model performance; and second, we employ a parameter-efficient cross-depth alignment technique, fine-tuning the pruned model to restore model accuracy. Our method achieves a 25% reduction in computational overhead on two TAD benchmarks (THUMOS14 and ActivityNet-1.3) to achieve lossless compression. More critically, we empirically show that our method is orthogonal to channel pruning methods and can be combined with it to yield further efficiency gains.
****
Differentiable Inverse Rendering with Interpretable Basis BRDFs
Hoon-Gyu Chung, Seokjun Choi, Seung-Hwan Baek; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 475-484
Inverse rendering seeks to reconstruct both geometry and spatially varying BRDFs (SVBRDFs) from captured images. To address the inherent ill-posedness of inverse rendering, basis BRDF representations are commonly used, modeling SVBRDFs as spatially varying blends of a set of basis BRDFs. However, existing methods often yield basis BRDFs that lack intuitive separation and have limited scalability to scenes of varying complexity. In this paper, we introduce a differentiable inverse rendering method that produces interpretable basis BRDFs. Our approach models a scene using 2D Gaussians, where the reflectance of each Gaussian is defined by a weighted blend of basis BRDFs. We efficiently render an image from the 2D Gaussians and basis BRDFs using differentiable rasterization and impose a specular-weighted rendering loss with the input flash photography images. During this analysis-by-synthesis optimization process of differentiable inverse rendering, we dynamically adjust the number of basis BRDFs to fit the target scene while encouraging sparsity in the basis weights. This ensures that the reflectance of each Gaussian is represented by only a few basis BRDFs. This approach enables the reconstruction of accurate geometry and interpretable basis BRDFs that are spatially separated. Consequently, the resulting scene representation, comprising basis BRDFs and 2D Gaussians, supports physically-based novel-view relighting and intuitive scene editing.
****
EquiPose: Exploiting Permutation Equivariance for Relative Camera Pose Estimation
Yuzhen Liu, Qiulei Dong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1127-1137
Relative camera pose estimation between two images is a fundamental task in 3D computer vision. Recently, many relative pose estimation networks have been explored for learning a mapping from two input images to their corresponding relative pose, however, the estimated relative poses by these methods do not have the intrinsic Pose Permutation Equivariance (PPE) property: the estimated relative pose from Image A to Image B should be the inverse of that from Image B to Image A. It means that permuting the input order of two images would cause these methods to obtain inconsistent relative poses. To address this problem, we firstly introduce the concept of PPE mapping, which indicates such a mapping that captures the intrinsic PPE property of relative poses. Then by enforcing the aforementioned PPE property, we propose a general framework for relative pose estimation, called EquiPose, which could easily accommodate various relative pose estimation networks in literature as its baseline models. We further theoretically prove that the proposed EquiPose framework could guarantee that its obtained mapping is a PPE mapping. Given a pre-trained baseline model, the proposed EquiPose framework could improve its performance even without fine-tuning, and could further boost its performance with fine-tuning. Experimental results on four public datasets

demonstrate that EquiPose could significantly improve the performances of variou
s state-of-the-art models.
****

Face Forgery Video Detection via Temporal Forgery Cue Unraveling
Zonghui Guo, Yingjie Liu, Jie Zhang, Haiyong Zheng, Shiguang Shan; Proceedings o
f the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7396-
7405
Face Forgery Video Detection (FFVD) is a critical yet challenging task in determ
ining whether a digital facial video is authentic or forged. Existing FFVD metho
ds typically focus on isolated spatial or coarsely fused spatiotemporal informat
ion, failing to leverage temporal forgery cues thus resulting in unsatisfactory
performance. We strive to unravel these cues across three progressive levels: mo
mentary anomaly, gradual inconsistency, and cumulative distortion. Accordingly,
we design a consecutive correlate module to capture momentary anomaly cues by co
rrelating interactions among consecutive frames. Then, we devise a future guide
module to unravel inconsistency cues by iteratively aggregating historical anoma
ly cues and gradually propagating them into future frames. Finally, we introduce
 a historical review module that unravels distortion cues via momentum accumulat
ion from future to historical frames. These three modules form our Temporal Forg
ery Cue Unraveling (TFCU) framework, sequentially highlighting spatial discrimin
ative features by unraveling temporal forgery cues bidirectionally between histo
rical and future frames. Extensive experiments and ablation studies demonstrate
the effectiveness of our TFCU method, achieving state-of-the-art performance acr
oss diverse unseen datasets and manipulation methods. Code is available at https
://github.com/zhenglab/TFCU.
****

Temporally Consistent Object-Centric Learning by Contrasting Slots
Anna Manasyan, Maximilian Seitzer, Filip Radovic, Georg Martius, Andrii Zadaianc
huk; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR
), 2025, pp. 5401-5411
Unsupervised object-centric learning from videos is a promising approach to extr
act structured representations from large, unlabeled collections of videos. To s
upport downstream tasks like autonomous control, these representations must be b
oth compositional and temporally consistent. Existing approaches based on recurr
ent processing often lack long-term stability across frames because their traini
ng objective does not enforce temporal consistency. In this work, we introduce a
 novel object-level temporal contrastive loss for video object-centric models th
at explicitly promotes temporal consistency. Our method significantly improves t
he temporal consistency of the learned object-centric representations, yielding
more reliable video decompositions that facilitate challenging downstream tasks
such as unsupervised object dynamics prediction. Furthermore, the inductive bias
 added by our loss strongly improves object discovery, leading to state-of-the-a
rt results on both synthetic and real-world datasets, outperforming even weakly-
supervised methods that leverage motion masks as additional cues.
****

MC^2: Multi-concept Guidance for Customized Multi-concept Generation
Jiaxiu Jiang, Yabo Zhang, Kailai Feng, Xiaohe Wu, Wenbo Li, Renjing Pei, Fan Li,
 Wangmeng Zuo; Proceedings of the Computer Vision and Pattern Recognition Confer
ence (CVPR), 2025, pp. 2802-2812
Customized text-to-image generation, which synthesizes images based on user-spec
ified concepts, has made significant progress in handling individual concepts. H
owever, when extended to multiple concepts, existing methods often struggle with
 properly integrating different models and avoiding the unintended blending of c
haracteristics from distinct concepts. In this paper, we propose MC^2, a novel a
pproach for multi-concept customization that enhances flexibility and fidelity t
hrough inference-time optimization. MC^2 enables the integration of multiple sin
gle-concept models with heterogeneous architectures. By adaptively refining atte
ntion weights between visual and textual tokens, our method ensures that image r
egions accurately correspond to their associated concepts while minimizing inter
ference between concepts. Extensive experiments demonstrate that MC^2 outperform

s training-based methods in terms of prompt-reference alignment. Furthermore, MC^2 can be seamlessly applied to text-to-image generation, providing robust compositional capabilities. To facilitate the evaluation of multi-concept customization, we also introduce a new benchmark, MC++. The code is available at https://github.com/JIANGJiaXiu/MC-2.

****

UniReal: Universal Image Generation and Editing via Learning Real-world Dynamics
Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, Hui Ding, Zhe Lin, Hengshuang Zhao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12501-12511

We introduce UniReal, a unified framework designed to address various image generation and editing tasks. Existing solutions often vary by tasks, yet share fundamental principles: preserving consistency between inputs and outputs while capturing visual variations. Inspired by recent video generation models that effectively balance consistency and variation across frames, we propose a unifying approach that treats image-level tasks as discontinuous video generation. Specifically, we treat varying numbers of input and output images as frames, enabling seamless support for tasks such as image generation, editing, composition, etc. Although designed for image-level tasks, we leverage videos as a scalable source for universal supervision. UniReal learns world dynamics from large-scale videos, demonstrating advanced capability in handling shadows, reflections, pose variation, and object interaction, while also exhibiting emergent capability for novel applications.

****

Pursuing Temporal-Consistent Video Virtual Try-On via Dynamic Pose Interaction
Dong Li, Wenqi Zhong, Wei Yu, Yingwei Pan, Dingwen Zhang, Ting Yao, Junwei Han, Tao Mei; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22648-22657

Video virtual try-on aims to seamlessly dress a subject in a video with a specific garment. The primary challenge involves preserving the visual authenticity of the garment while dynamically adapting to the pose and physique of the subject. While existing methods have predominantly focused on image-based virtual try-on, extending these techniques directly to videos often results in temporal inconsistencies. Most current video virtual try-on approaches alleviate this challenge by incorporating temporal modules, yet still overlook the critical spatiotemporal pose interactions between human and garment. Effective pose interactions in videos should not only consider spatial alignment between human and garment poses in each frame but also account for the temporal dynamics of human poses throughout the entire video. With such motivation, we propose a new framework, namely Dynamic Pose Interaction Diffusion Models (DPIDM), to leverage diffusion models to delve into dynamic pose interactions for video virtual try-on. Technically, DPIDM introduces a skeleton-based pose adapter to integrate synchronized human and garment poses into the denoising network. A hierarchical attention module is then exquisitely designed to model intra-frame human-garment pose interactions and long-term human pose dynamics across frames through pose-aware spatial and temporal attention mechanisms. Moreover, DPIDM capitalizes on a temporal regularized attention loss between consecutive frames to enhance temporal consistency. Extensive experiments conducted on VITON-HD, VVT and ViViD datasets demonstrate the superiority of our DPIDM against the baseline methods. Notably, DPIDM achieves VFID score of 0.506 on VVT dataset, leading to 60.5% improvement over the state-of-the-art GPD-VVTO approach.

****

Exploring Contextual Attribute Density in Referring Expression Counting
Zhicheng Wang, Zhiyu Pan, Zhan Peng, Jian Cheng, Liwen Xiao, Wei Jiang, Zhiguo Cao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19587-19596

Referring expression counting (REC) algorithms are for more flexible and interactive counting ability across varied fine-grained text expressions. However, the requirement for fine-grained attribute understanding poses challenges for prior

arts, as they struggle to accurately align attribute information with correct visual patterns. Given the proven importance of "visual density", it is presumed that the limitations of current REC approaches stem from an under-exploration of "contextual attribute density" (CAD). In the scope of REC, we define the CAD as the measure of the information intensity of one certain fine-grained attribute in visual regions. To model the the CAD, we propose a U-shape CAD estimator in which referring expression and multi-scale visual features from GroundingDINO can interact with each other. With additional density supervisions, we can effectively encode CAD, which is subsequently decoded via a novel attention procedure with CAD-refined queries. Integrating all these contributions, our framework significantly outperforms state-of-the-art REC methods, achieves 30% error reduction in counting metics and a 10% improvement in localization accuracy. The surprising results shed lights on the significance of contextual attribute density for REC. Code will be at github.com/Xu3XiWang/CAD-GD.

****

DINOv2 Meets Text: A Unified Framework for Image- and Pixel-Level Vision-Language Alignment

Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, Oriane Siméoni, Huy V. Vo, Patrick Labatut, Piotr Bojanowski; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24905-24916

Self-supervised visual foundation models produce powerful embeddings that achieve remarkable performance on a wide range of downstream tasks. However, unlike vision-language models such as CLIP, self-supervised visual features are not readily aligned with language, hindering their adoption in open-vocabulary tasks. Our method unlocks this new ability for DINOv2, a widely used self-supervised visual encoder. We build upon the LiT training strategy, which trains a text encoder to align with a frozen vision model but leads to unsatisfactory results on dense tasks. We propose several key ingredients to improve performance on both global and dense tasks, such as concatenating the [CLS] token with the patch average to train the alignment and curating data using both text and image modalities. With these, we successfully train a CLIP-like model with only a fraction of the computational cost compared to CLIP while achieving state-of-the-art results in zero-shot classification and open-vocabulary semantic segmentation.

****

Learning Affine Correspondences by Integrating Geometric Constraints

Pengju Sun, Banglei Guan, Zhenbao Yu, Yang Shang, Qifeng Yu, Daniel Barath; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27038-27048

Affine correspondences have received significant attention due to their benefits in tasks like image matching and pose estimation. Existing methods for extracting affine correspondences still have many limitations in terms of performance; thus, exploring a new paradigm is crucial. In this paper, we present a new pipeline designed for extracting accurate affine correspondences by integrating dense matching and geometric constraints. Specifically, a novel extraction framework is introduced, with the aid of dense matching and a novel keypoint scale and orientation estimator. For this purpose, we propose loss functions based on geometric constraints, which can effectively improve accuracy by supervising neural networks to learn feature geometry. The experimental show that the accuracy and robustness of our method outperform the existing ones in image matching tasks. To further demonstrate the effectiveness of the proposed method, we applied it to relative pose estimation. Affine correspondences extracted by our method lead to more accurate poses than the baselines on a range of real-world datasets. The source code will be made public.

****

UCOD-DPL: Unsupervised Camouflaged Object Detection via Dynamic Pseudo-label Learning

Weiqi Yan, Lvhai Chen, Huaijia Kou, Shengchuan Zhang, Yan Zhang, Liujuan Cao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30365-30375

Unsupervised Camoflaged Object Detection (UCOD) has gained attention since it doesn't need to rely on extensive pixel-level labels. Existing UCOD methods typically generate pseudo-labels using fixed strategies and train 1 x1 convolutional layers as a simple decoder, leading to low performance compared to fully-supervised methods. We emphasize two drawbacks in these approaches: 1). The model is prone to fitting incorrect knowledge due to the pseudo-label containing substantial noise. 2). The simple decoder fails to capture and learn the semantic features of camouflaged objects, especially for small-sized objects, due to the low-resolution pseudo-labels and severe confusion between foreground and background pixels. To this end, we propose a UCOD method with a teacher-student framework via Dynamic Pseudo-label Learning called UCOD-DPL, which contains an Adaptive Pseudo-label Module (APM), a Dual-Branch Adversarial (DBA) decoder, and a Look-Twice mechanism. The APM module adaptively combines pseudo-labels generated by fixed strategies and the teacher model to prevent the model from overfitting incorrect knowledge while preserving the ability for self-correction; the DBA decoder takes adversarial learning of different segmentation objectives, guides the model to overcome the foreground-background confusion of camouflaged objects, and the Look-Twice mechanism mimics the human tendency to zoom in on camouflaged objects and performs secondary refinement on small-sized objects. Extensive experiments show that our method demonstrates outstanding performance, even surpassing some existing fully supervised methods. Our code will be released soon.

****

Geometry in Style: 3D Stylization via Surface Normal Deformation

Nam Anh Dinh, Itai Lang, Hyunwoo Kim, Oded Stein, Rana Hanocka; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28456-28467

We present Geometry in Style, a new method for identity-preserving mesh stylization. Existing techniques either adhere to the original shape through overly restrictive deformations such as bump maps or significantly modify the input shape using expressive deformations that may introduce artifacts or alter the identity of the source shape. In contrast, we represent a deformation of a triangle mesh as a target normal vector for each vertex neighborhood. The deformations we recover from target normals are expressive enough to enable detailed stylizations yet restrictive enough to preserve the shape's identity. We achieve such deformations using our novel differentiable As-Rigid-As-Possible (dARAP) layer, a neural-network-ready adaptation of the classical ARAP algorithm which we use to solve for per-vertex rotations and deformed vertices. As a differentiable layer, dARAP is paired with a visual loss from a text-to-image model to drive deformations toward style prompts, altogether giving us Geometry in Style.

****

Multi-modal Vision Pre-training for Medical Image Analysis

Shaohao Rui, Lingzhi Chen, Zhenyu Tang, Lilong Wang, Mianxin Liu, Shaoting Zhang, Xiaosong Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5164-5174

Self-supervised learning has greatly facilitated medical image analysis by suppressing the training data requirement for real-world applications. Current paradigms predominantly rely on self-supervision within uni-modal image data, thereby neglecting the inter-modal correlations essential for effective learning of cross-modal image representations. This limitation is particularly significant for naturally grouped multi-modal data, e.g., multi-parametric MRI scans for a patient undergoing various functional imaging protocols in the same study. To bridge this gap, we conduct a novel multi-modal image pre-training with three proxy tasks to facilitate the learning of cross-modality representations and correlations using multi-modal brain MRI scans (over 2.4 million images in 16,022 scans of 3,755 patients), i.e., cross-modal image reconstruction, modality-aware contrastive learning, and modality template distillation. To demonstrate the generalizability of our pre-trained model, we conduct extensive experiments on various benchmarks with ten downstream tasks. The superior performance of our method is reported in comparison to state-of-the-art pre-training methods, with Dice Score improvement of 0.28%-14.47% across six segmentation benchmarks and a consistent accur

acy boost of 0.65%-18.07% in four individual image classification tasks.
****

SegMAN: Omni-scale Context Modeling with State Space Models and Local Attention for Semantic Segmentation

Yunxiang Fu, Meng Lou, Yizhou Yu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19077-19087

High-quality semantic segmentation relies on three key capabilities: global context modeling, local detail encoding, and multi-scale feature extraction. However, recent methods struggle to possess all these capabilities simultaneously. Hence, we aim to empower segmentation networks to simultaneously carry out efficient global context modeling, high-quality local detail encoding, and rich multi-scale feature representation for varying input resolutions. In this paper, we introduce SegMAN, a novel linear-time model comprising a hybrid feature encoder dubbed SegMAN Encoder, and a decoder based on state space models. Specifically, the SegMAN Encoder synergistically integrates sliding local attention with dynamic state space models, enabling highly efficient global context modeling while preserving fine-grained local details. Meanwhile, the MMSCopE module in our decoder enhances multi-scale context feature extraction and adaptively scales with the input resolution. Our SegMAN-B Encoder achieves 85.1% ImageNet-1k accuracy (+1.5% over VMamba-S with fewer parameters). When paired with our decoder, the full SegMAN-B model achieves 52.6% mIoU on ADE20K (+1.6% over SegNeXt-L with 15% fewer GFLOPs), 83.8% mIoU on Cityscapes (+2.1% over SegFormer-B3 with half the GFLOPs), and 1.6% higher mIoU than VWFormer-B3 on COCO-Stuff with lower GFLOPs. Our code is available at https://github.com/yunxiangfu2001/SegMAN.
****

STEP: Enhancing Video-LLMs' Compositional Reasoning by Spatio-Temporal Graph-guided Self-Training

Haiyi Qiu, Minghe Gao, Long Qian, Kaihang Pan, Qifan Yu, Juncheng Li, Wenjie Wang, Siliang Tang, Yueting Zhuang, Tat-Seng Chua; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3284-3294

Video Large Language Models (Video-LLMs) have recently shown strong performance in basic video understanding tasks, such as captioning and coarse-grained question answering, but struggle with compositional reasoning that requires multi-step spatio-temporal inference across object relations, interactions, and events. The hurdles to enhancing this capability include extensive manual labor, the lack of spatio-temporal compositionality in existing training data and the absence of explicit reasoning supervision. In this paper, we propose STEP, a novel graph-guided self-training method that enables Video-LLMs to generate reasoning-rich fine-tuning data from any raw videos to improve itself. Specifically, we first induce Spatio-Temporal Scene Graph (STSG) representation of diverse videos to capture fine-grained, multi-granular video semantics. Then, the STSGs guide the derivation of multi-step reasoning Question-Answer (QA) data with Chain-of-Thought (CoT) rationales. Both answers and rationales are integrated as training objective, aiming to enhance model's reasoning abilities by supervision over explicit reasoning steps. Experimental results demonstrate the effectiveness of STEP across models of varying scales, with a significant 21.3% improvement in tasks requiring three or more reasoning steps. Furthermore, it achieves superior performance with a minimal amount of self-generated rationale-enriched training samples in both compositional reasoning and comprehensive understanding benchmarks, highlighting the broad applicability and vast potential.
****

OmniFlow: Any-to-Any Generation with Multi-Modal Rectified Flows

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Zichun Liao, Yusuke Kato, Kazuki Kozuka, Aditya Grover; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13178-13188

We introduce OminiFlow, a novel generative model designed for any-to-any generation tasks such as text-to-image, text-to-audio, and audio-to-image synthesis. OminiFlow advances the rectified flow (RF) framework used in text-to-image models to handle the joint distribution of multiple modalities. It outperforms previous any-to-any models on a wide range of tasks, such as text-to-image and text-to-a

udio synthesis. Our work offers three key contributions: First, we extend RF to a multi-modal setting and introduce a novel guidance mechanism, enabling users to flexibly control the alignment between different modalities in the generated outputs. Second, we propose a novel architecture that extends the text-to-image MMDiT architecture of Stable Diffusion 3 and enables audio and text generation. The extended modules can be efficiently pretrained individually and merged with the vanilla text-to-image MMDiT for fine-tuning. Lastly, we conduct a comprehensive study on the design choices of rectified flow transformers for large-scale audio and text generation, providing valuable insights into optimizing performance across diverse modalities.
****

PVC: Progressive Visual Token Compression for Unified Image and Video Processing in Large Vision-Language Models
Chenyu Yang, Xuan Dong, Xizhou Zhu, Weijie Su, Jiahao Wang, Hao Tian, Zhe Chen, Wenhai Wang, Lewei Lu, Jifeng Dai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24939-24949
Large Vision-Language Models (VLMs) have been extended to understand both images and videos. Visual token compression is leveraged to reduce the considerable token length of visual inputs. To meet the needs of different tasks, existing high-performance models usually process images and videos separately with different token compression strategies, limiting the capabilities of combining images and videos. To this end, we extend each image into a "static" video and introduce a unified token compression strategy called Progressive Visual Token Compression (PVC), where the tokens of each frame are progressively encoded and adaptively compressed to supplement the information not extracted from previous frames. Video tokens are efficiently compressed with exploiting the inherent temporal redundancy. Images are repeated as static videos, and the spatial details can be gradually supplemented in multiple frames. PVC unifies the token compressing of images and videos. With a limited number of tokens per frame (64 tokens by default), spatial details and temporal changes can still be preserved. Experiments show that our model achieves state-of-the-art performance across various video understanding benchmarks, including long video tasks and fine-grained short video tasks. Meanwhile, our unified token compression strategy incurs no performance loss on image benchmarks, particularly in detail-sensitive tasks.
****

LIM: Large Interpolator Model for Dynamic Reconstruction
Remy Sabathier, Niloy J. Mitra, David Novotny; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6154-6164
Reconstructing dynamic assets from video data is central to many in computer vision and graphics tasks. Existing 4D reconstruction approaches are limited by category-specific models or slow optimization-based methods. Inspired by the recent Large Reconstruction Model (LRM), we present the Large Interpolation Model (LIM), a transformer-based feed-forward solution, guided by a novel causal consistency loss, for interpolating implicit 3D representations across time. Given implicit 3D representations at times t_0 and t_1, LIM produces a deformed shape at any continuous time t\in[t_0,t_1] delivering high-quality interpolations in seconds (per frame).Furthermore, LIM allows explicit mesh tracking across time, producing a consistently uv-textured mesh sequence ready for integration into existing production pipelines. We also use LIM, in conjunction with a diffusion-based multiview generator, to produce dynamic 4D reconstructions from monocular videos. We evaluate LIM on various dynamic datasets, benchmarking against image-space interpolation methods (e.g., FiLM) and direct triplane linear interpolation, and demonstrate clear advantages. In summary, LIM is the first feed-forward model capable of high-speed tracked 4D asset reconstruction across diverse categories.
****

Multiple Object Tracking as ID Prediction
Ruopeng Gao, Ji Qi, Limin Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27883-27893
Multi-Object Tracking (MOT) has been a long-standing challenge in video understanding. A natural and intuitive approach is to split this task into two parts: ob

ject detection and association. Most mainstream methods employ meticulously crafted heuristic techniques to maintain trajectory information and compute cost matrices for object matching. Although these methods can achieve notable tracking performance, they often require a series of elaborate handcrafted modifications while facing complicated scenarios. We believe that manually assumed priors limit the method's adaptability and flexibility in learning optimal tracking capabilities from domain-specific data. Therefore, we introduce a new perspective that treats Multiple Object Tracking as an in-context ID Prediction task, transforming the aforementioned object association into an end-to-end trainable task. Based on this, we propose a simple yet effective method termed MOTIP. Given a set of trajectories carried with ID information, MOTIP directly decodes the ID labels for current detections to accomplish the association process. Without using tailored or sophisticated architectures, our method achieves state-of-the-art results across multiple benchmarks by solely leveraging object-level features as tracking cues. The simplicity and impressive results of MOTIP leave substantial room for future advancements, thereby making it a promising baseline for subsequent research. Our code and checkpoints are released at https://github.com/MCG-NJU/MOTIP.

****

AutoPresent: Designing Structured Visuals from Scratch

Jiaxin Ge, Zora Zhiruo Wang, Xuhui Zhou, Yi-Hao Peng, Sanjay Subramanian, Qinyue Tan, Maarten Sap, Alane Suhr, Daniel Fried, Graham Neubig, Trevor Darrell; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2902-2911

Designing structured visuals such as presentation slides is essential for communicative needs, necessitating both content creation and visual planning skills. In this work, we tackle the challenge of automated slide generation, where models produce slide presentations from natural language (NL) instructions. We first introduce the SlidesBench benchmark, the first benchmark for slide generation with 7k training and 585 testing examples derived from 310 slide decks across 10 domains. SlidesBench supports evaluations that are (i) reference-based to measure similarity to a target slide, and (ii) reference-free to measure the design quality of generated slides alone. We benchmark end-to-end image generation and program generation methods with a variety of models, and find that programmatic methods produce higher-quality slides in user-interactable formats. Built on the success of program generation, we create AutoPresent, an 8B Llama-based model trained on 7k pairs of instructions paired with code for slide generation, and achieve results comparable to the closed-source model GPT-4o. We further explore iterative design refinement where the model is tasked to self-refine its own output, and we found that this process improves the slide's quality. We hope that our work will provide a basis for future work on generating structured visuals.

****

SLAM3R: Real-Time Dense Scene Reconstruction from Monocular RGB Videos

Yuzheng Liu, Siyan Dong, Shuzhe Wang, Yingda Yin, Yanchao Yang, Qingnan Fan, Baoquan Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16651-16662

In this paper, we introduce SLAM3R, a novel and effective system for real-time, high-quality, dense 3D reconstruction using RGB videos. SLAM3R provides an end-to-end solution by seamlessly integrating local 3D reconstruction and global coordinate registration through feed-forward neural networks. Given an input video, the system first converts it into overlapping clips using a sliding window mechanism. Unlike traditional pose optimization-based methods, SLAM3R directly regresses 3D pointmaps from RGB images in each window and progressively aligns and deforms these local pointmaps to create a globally consistent scene reconstruction - all without explicitly solving any camera parameters. Experiments across datasets consistently show that SLAM3R achieves state-of-the-art reconstruction accuracy and completeness while maintaining real-time performance at 20+ FPS. Code available at: https://github.com/PKU-VCL-3DV/SLAM3R.

****

PIDLoc: Cross-View Pose Optimization Network Inspired by PID Controllers

Wooju Lee, Juhye Park, Dasol Hong, Changki Sung, Youngwoo Seo, DongWan Kang, Hyu n Myung; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21981-21990
Accurate localization is essential for autonomous driving, but GNSS-based method s struggle in challenging environments such as urban canyons. Cross-view pose op timization offers an effective solution for localization by directly estimating vehicle pose using satellite-view images. However, existing methods primarily re ly on cross-view features at a given pose, neglecting fine-grained contexts for precision and global contexts for robustness against large initial pose errors. To overcome these limitations, we propose PIDLoc, a novel cross-view pose optimi zation approach inspired by the proportional-integral-derivative (PID) controlle r. Using RGB images and LiDAR data, the PIDLoc models cross-view feature relatio nships through the PID branches and estimates pose via the spatially aware pose estimator (SPE). To enhance localization accuracy, the PID branches leverage fea ture differences for local context (P), aggregated feature differences for globa l context (I), and gradients of feature differences for fine-grained context (D) . Integrated with the PID branches, the SPE captures spatial relationships withi n the PID-branch features for consistent localization. Experimental results demo nstrate that the PIDLoc achieves state-of-the-art performance in cross-view pose estimation for the KITTI dataset, reducing position error by 37.8% compared wit h the previous state-of-the-art. Our code is available at https://github.com/url -kaist/PIDLoc.
****

VisionArena: 230k Real World User-VLM Conversations with Preference Labels
Christopher Chou, Lisa Dunlap, Koki Mashita, Krishna Mandal, Trevor Darrell, Ion Stoica, Joseph E. Gonzalez, Wei-Lin Chiang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3877-3887
The growing adoption and capabilities of vision-language models (VLMs) demand be nchmarks that reflect real-world user interactions. We introduce VisionArena, th e largest existing dataset of crowdsourced real-world conversations between user s and VLMs. While most visual question-answering datasets focus on close-ended p roblems or synthetic scenarios, VisionArena contains a wide variety of closed an d open ended problems across 230K conversations, 73K unique users, 138 languages , and 45 VLMs. VisionArena consists of VisionArena-Chat, a set of 200k single-tu rn and multi-turn chat logs with a VLM, VisionArena-Battle, a set of 30K conver sations between a user and 2 anonymous VLMs with preference votes, and VisionAre na-Bench, an automatic benchmark consisting of 500 diverse user prompts which ca n be used to cheaply approximate model rankings.We analyze these datasets and hi ghlight the types of question asked by users, the influence of style on user pre ference, and areas where models often fall short. We find that more open-ended q uestions like captioning and humor are heavily influenced by style, which causes certain models like Reka Flash and InternVL which are tuned for style to perfor m significantly better on these categories compared to other categories. We show that VisionArena-Chat and VisionArena-Battle can be used for post-training to a lign VLMs to human preferences through supervised fine-tuning. Compared to the p opular instruction tuning dataset Llava-Instruct-158K, finetuning the same base model on VisionArena results in a a 17 point improvement on MMMU and a 46 point improvement on the WildVision human preference benchmark. Lastly, we show that r unning automatic VLM evaluation on VisionArena-Bench results in a model ranking which is largely consistent with major online preference benchmarks. We release both VisionArena and our finetuned model to further VLM development.
****

FAM Diffusion: Frequency and Attention Modulation for High-Resolution Image Gene ration with Stable Diffusion
Haosen Yang, Adrian Bulat, Isma Hadji, Hai X. Pham, Xiatian Zhu, Georgios Tzimir opoulos, Brais Martinez; Proceedings of the Computer Vision and Pattern Recognit ion Conference (CVPR), 2025, pp. 2459-2468
Diffusion models are proficient at generating high-quality images. They are howe ver effective only when operating at the resolution used during training. Infere nce at a scaled resolution leads to repetitive patterns and structural distortio

ns. Retraining at higher resolutions quickly becomes prohibitive. Thus, methods enabling pre-existing diffusion models to operate at flexible test-time resolut ions are highly desirable. Previous works suffer from frequent artifacts and oft en introduce large latency overheads. We propose two simple modules that combin e to solve these issues. We introduce a Frequency Modulation (FM) module that le verages the Fourier domain to improve the global structure consistency, and an A ttention Modulation (AM) module which improves the consistency of local texture patterns, a problem largely ignored in prior works. Our method, coined FAM diffu sion, can seamlessly integrate into any latent diffusion model and requires no a dditional training. Extensive qualitative results highlight the effectiveness o f our method in addressing structural and local artifacts, while quantitative re sults show state-of-the-art performance. Also, our method avoids redundant infer ence tricks for improved consistency such as patch-based or progressive generati on, leading to negligible latency overheads.
****

DreamOmni: Unified Image Generation and Editing
Bin Xia, Yuechen Zhang, Jingyao Li, Chengyao Wang, Yitong Wang, Xinglong Wu, Bei Yu, Jiaya Jia; Proceedings of the Computer Vision and Pattern Recognition Confe rence (CVPR), 2025, pp. 28533-28543
Currently, the success of large language models (LLMs) illustrates that a unifie d multitasking approach can significantly enhance model usability, streamline de ployment, and foster synergistic benefits across different tasks. However, in co mputer vision, while text-to-image (T2I) models have significantly improved gene ration quality through scaling up, their framework design did not initially cons ider how to unify with downstream tasks, such as various types of editing. To ad dress this, we introduce DreamOmni, a unified model for image generation and edi ting. We begin by analyzing existing frameworks and the requirements of downstre am tasks, proposing a unified framework that integrates both T2I models and vari ous editing tasks. Furthermore, another key challenge is the efficient creation of high-quality editing data, particularly for instruction-based and drag-based editing. To this end, we develop a synthetic data pipeline using sticker-like el ements to synthesize accurate, high-quality datasets efficiently, which enables editing data scaling up for unified model training. For training, DreamOmni joi ntly trains T2I generation and downstream tasks. T2I training enhances the model 's understanding of specific concepts and improves generation quality, while edi ting training helps the model grasp the nuances of the editing task. This collab oration significantly boosts editing performance. Extensive experiments confirm the effectiveness of DreamOmni. The code and model will be released.
****

Hash3D: Training-free Acceleration for 3D Generation
Xingyi Yang, Songhua Liu, Xinchao Wang; Proceedings of the Computer Vision and P attern Recognition Conference (CVPR), 2025, pp. 21481-21491
The quality of 3D generative modeling has been notably improved by the adoption of 2D diffusion models. Despite this progress, the cumbersome optimization proce ss per sepresents a critical problem to efficiency. In this paper, we introduce Hash3D, a universal acceleration for 3D score distillation sampling (SDS) witho ut model training.Central to Hash3D is the observation that images rendered from similar camera positions and diffusion time-steps often have redundant feature maps. By hashing and reusing these feature maps across nearby timesteps and came ra angles, Hash3D eliminates unnecessary calculations. We implement this through an adaptive grid-based hashing. As a result, it largely speeds up the process o f 3D generation. Surprisingly, this feature-sharing mechanism not only makes gen eration faster but also improves the smoothness and view consistency of the synt hesized 3D objects. Our experiments covering 5 text-to-3D and 3 image-to-3D mode ls, demonstrate Hash3D's versatility to speed up optimization, enhancing efficie ncy by 1.5~ 4x. Additionally, Hash3D's integration with 3D Gaussian splatting la rgely speeds up 3D model creation, reducing text-to-3D conversion to about 10 mi nutes and image-to-3D conversion to 30 seconds. The project page is https://adam dad.github.io/hash3D/.
****

SemGeoMo: Dynamic Contextual Human Motion Generation with Semantic and Geometric Guidance
Peishan Cong, Ziyi Wang, Yuexin Ma, Xiangyu Yue; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17561-17570
Generating reasonable and high-quality human interactive motions in a given dynamic environment is crucial for understanding, modeling, transferring, and applying human behaviors to both virtual and physical robots. In this paper, we introduce an effective method, SemGeoMo, for dynamic contextual human motion generation, which fully leverages the text-affordance-joint multi-level semantic and geometric guidance in the generation process, improving the semantic rationality and geometric correctness of generative motions. Our method achieves state-of-the-art performance on three datasets and demonstrates superior generalization capability for diverse interaction scenarios.
****
MultiGO: Towards Multi-level Geometry Learning for Monocular 3D Textured Human Reconstruction
Gangjian Zhang, Nanjie Yao, Shunsi Zhang, Hanfeng Zhao, Guoliang Pang, Jian Shu, Hao Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 338-347
This paper investigates the research task of reconstructing the 3D clothed human body from a monocular image. Due to the inherent ambiguity of single-view input, existing approaches leverage pre-trained SMPL(-X) estimation models or generative models to provide auxiliary information for human reconstruction. However, these methods capture only the general human body geometry and overlook specific geometric details, leading to inaccurate skeleton reconstruction, incorrect joint positions, and unclear cloth wrinkles. In response to these issues, we propose a multi-level geometry learning framework. Technically, we design three key components: skeleton-level enhancement, joint-level augmentation, and wrinkle-level refinement modules. Specifically, we effectively integrate the projected 3D Fourier features into a Gaussian reconstruction model, introduce perturbations to improve joint depth estimation during training, and refine the human coarse wrinkles by resembling the de-noising process of the diffusion model. Extensive quantitative and qualitative experiments on two test sets show the superior performance of our approach compared to state-of-the-art (SOTA) methods.
****
Generative Photomontage
Sean J. Liu, Nupur Kumari, Ariel Shamir, Jun-Yan Zhu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7931-7941
Text-to-image models are powerful tools for image creation. However, the generation process is akin to a dice roll and makes it difficult to achieve a single image that captures everything a user wants. In this paper, we propose a framework for creating the desired image by compositing it from various parts of generated images, in essence forming a Generative Photomontage. Given a stack of images generated by ControlNet using the same input condition and different seeds, we let users select desired parts from the generated results using a brush stroke interface. We introduce a novel technique that takes in the user's brush strokes, segments the generated images using a graph-based optimization in diffusion feature space, and then composites the segmented regions via a new feature-space blending method. Our method faithfully preserves the user-selected regions while compositing them harmoniously. We demonstrate that our flexible framework can be used for many applications, including generating new appearance combinations, fixing incorrect shapes and artifacts, and improving prompt alignment. We show compelling results for each application and demonstrate that our method outperforms existing image blending methods and various baselines.
****
Multi-view Reconstruction via SfM-guided Monocular Depth Estimation
Haoyu Guo, He Zhu, Sida Peng, Haotong Lin, Yunzhi Yan, Tao Xie, Wenguan Wang, Xiaowei Zhou, Hujun Bao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5272-5282
This paper aims to reconstruct the scene geometry from multi-view images with st

rong robustness and high quality. Previous learning-based methods incorporate ne
ural networks into the multi-view stereo matching and have shown impressive reco
nstruction results. However, due to the reliance on matching across input images
, they typically suffer from high GPU memory consumption and tend to fail in spa
rse view scenarios. To overcome this problem, we develop a new pipeline, named M
urre, for multi-view geometry reconstruction of 3D scenes based on SfM-guided mo
nocular depth estimation. For input images, Murre first recover the SfM point cl
oud that captures the global scene structure, and then use it to guide a conditi
onal diffusion model to produce multi-view metric depth maps for the final TSDF
fusion. By predicting the depth map from a single image, Murre bypasses the mult
i-view matching step and naturally resolves the issues of previous MVS-based met
hods. In addition, the diffusion-based model can easily leverage the powerful pr
iors of 2D foundation models, achieving good generalization ability across diver
se real-world scenes. To obtain multi-view consistent depth maps, our key design
is providing effective guidance on the diffusion model through the SfM point cl
oud, which is a condensed form of multi-view information, highlighting the scene
's salient structure, and can be readily transformed into point maps to drive th
e image-space estimation process. We evaluate the reconstruction quality of Murr
e in various types of real-world datasets including indoor, streetscapes, and ae
rial scenes, surpassing state-of-the-art MVS-based and implicit neural reconstru
ction-based methods. The code and supplementary materials are available at https
://zju3dv.github.io/murre/.
****

Learning Hazing to Dehazing: Towards Realistic Haze Generation for Real-World Im
age Dehazing
Ruiyi Wang, Yushuo Zheng, Zicheng Zhang, Chunyi Li, Shuaicheng Liu, Guangtao Zha
i, Xiaohong Liu; Proceedings of the Computer Vision and Pattern Recognition Conf
erence (CVPR), 2025, pp. 23091-23100
Existing real-world image dehazing methods primarily attempt to fine-tune pre-tr
ained models or adapt their inference procedures, thus heavily relying on the pr
e-trained models and associated training data. Moreover, restoring heavily disto
rted information under dense haze requires generative diffusion models, whose po
tential in dehazing remains underutilized partly due to their lengthy sampling p
rocesses. To address these limitations, we introduce a novel hazing-dehazing pip
eline consisting of a Realistic Hazy Image Generation framework (HazeGen) and a
Diffusion-based Dehazing framework (DiffDehaze). Specifically, HazeGen harnesses
robust generative diffusion priors of real-world hazy images embedded in a pre-
trained text-to-image diffusion model. By employing specialized hybrid training
and blended sampling strategies, HazeGen produces realistic and diverse hazy ima
ges as high-quality training data for DiffDehaze. To alleviate the inefficiency
and fidelity concerns associated with diffusion-based methods, DiffDehaze adopts
an Accelerated Fidelity-Preserving Sampling process (AccSamp). The core of AccS
amp is the Tiled Statistical Alignment Operation (AlignOp), which can provide a
clean and faithful dehazing estimate within a small fraction of sampling steps t
o reduce complexity and enable effective fidelity guidance. Extensive experiment
s demonstrate the superior dehazing performance and visual quality of our approa
ch over existing methods. The code is available at https://github.com/ruiyi-w/Le
arning-Hazing-to-Dehazing.
****

HuMoCon: Concept Discovery for Human Motion Understanding
Qihang Fang, Chengcheng Tang, Bugra Tekin, Shugao Ma, Yanchao Yang; Proceedings
of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7179
-7190
We present HuMoCon, a novel motion-video understanding framework designed for ad
vanced human behavior analysis. The core of our method is a human motion concept
discovery framework that efficiently trains multi-modal encoders to extract sem
antically meaningful and generalizable features. HuMoCon addresses key challenge
s in motion concept discovery for understanding and reasoning, including the lac
k of explicit multi-modality feature alignment and the loss of high-frequency in
formation in masked autoencoding frameworks. Our approach integrates a feature a

lignment strategy that leverages video for contextual understanding and motion f
or fine-grained interaction modeling, further with a velocity reconstruction mec
hanism to enhance high-frequency feature expression and mitigate temporal over-s
moothing. Comprehensive experiments on standard benchmarks demonstrate that HuMo
Con enables effective motion concept discovery and significantly outperforms sta
te-of-the-art methods in training large models for human motion understanding.
****

RUBIK: A Structured Benchmark for Image Matching across Geometric Challenges
Thibaut Loiseau, Guillaume Bourmaud; Proceedings of the Computer Vision and Patt
ern Recognition Conference (CVPR), 2025, pp. 27070-27080
Camera pose estimation is crucial for many computer vision applications, yet exi
sting benchmarks offer limited insight into method limitations across different
geometric challenges. We introduce RUBIK, a novel benchmark that systematically
evaluates image matching methods across well-defined geometric difficulty levels
. Using three complementary criteria - overlap, scale ratio, and viewpoint angle
 - we organize 16.5K image pairs from nuScenes into 33 difficulty levels. Our co
mprehensive evaluation of 14 methods reveals that while recent detector-free app
roaches achieve the best performance (>47% success rate), they come with signifi
cant computational overhead compared to detector-based methods (150-600ms vs. 40
-70ms). Even the best performing method succeeds on only 54.8% of the pairs, hig
hlighting substantial room for improvement, particularly in challenging scenario
s combining low overlap, large scale differences, and extreme viewpoint changes.
 Benchmark will be made publicly available.
****

Fast and Accurate Gigapixel Pathological Image Classification with Hierarchical
Distillation Multi-Instance Learning
Jiuyang Dong, Junjun Jiang, Kui Jiang, Jiahan Li, Yongbing Zhang; Proceedings of
 the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30818-
30828
Although multi-instance learning (MIL) has succeeded in pathological image class
ification, it faces the challenge of high inference costs due to processing nume
rous patches from gigapixel whole slide images (WSIs).To address this, we propos
e HDMIL, a hierarchical distillation multi-instance learning framework that achi
eves fast and accurate classification by eliminating irrelevant patches.HDMIL co
nsists of two key components: the dynamic multi-instance network (DMIN) and the
lightweight instance pre-screening network (LIPN). DMIN operates on high-resolut
ion WSIs, while LIPN operates on the corresponding low-resolution counterparts.D
uring training, DMIN are trained for WSI classification while generating attenti
on-score-based masks that indicate irrelevant patches.These masks then guide the
 training of LIPN to predict the relevance of each low-resolution patch.During t
esting, LIPN first determines the useful regions within low-resolution WSIs, whi
ch indirectly enables us to eliminate irrelevant regions in high-resolution WSIs
, thereby reducing inference time without causing performance degradation.In add
ition, we further design the first Chebyshev-polynomials-based Kolmogorov-Arnold
 classifier in computational pathology, which enhances the performance of HDMIL
through learnable activation layers.Extensive experiments on three public datase
ts demonstrate that HDMIL outperforms previous state-of-the-art methods, e.g., a
chieving improvements of 3.13% in AUC while reducing inference time by 28.6% on
the Camelyon16 dataset.The project will be available.
****

FreeScene: Mixed Graph Diffusion for 3D Scene Synthesis from Free Prompts
Tongyuan Bai, Wangyuanfan Bai, Dong Chen, Tieru Wu, Manyi Li, Rui Ma; Proceeding
s of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 58
93-5903
Controllability plays a crucial role in the practical applications of 3D indoor
scene synthesis. Existing works either allow rough language-based control, that
is convenient but lacks fine-grained scene customization, or employ graph-based
control, which offers better controllability but demands considerable knowledge
for the cumbersome graph design process. To address these challenges, we present
 FreeScene, a user-friendly framework that enables both convenient and effective

control for indoor scene synthesis. Specifically, FreeScene supports free-form user inputs including text description and/or reference images, allowing users to express versatile design intentions. The user inputs are adequately analyzed and integrated into a graph representation by a VLM-based Graph Designer. We then propose MG-DiT, a Mixed Graph Diffusion Transformer, which performs graph-aware denoising to enhance scene generation. Our MG-DiT not only excels at preserving graph structure but also offers broad applicability to various tasks, including, but not limited to, text-to-scene, graph-to-scene, and rearrangement, all within a single model. Extensive experiments demonstrate that FreeScene provides an efficient and user-friendly solution that unifies text-based and graph-based scene synthesis, outperforming state-of-the-art methods in terms of both generation quality and controllability in a range of applications.

****

Rethinking Correspondence-based Category-Level Object Pose Estimation

Huan Ren, Wenfei Yang, Shifeng Zhang, Tianzhu Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1170-1179

Category-level object pose estimation aims to determine the pose and size of arbitrary objects within given categories. Existing two-stage correspondence-based methods first establish correspondences between camera and object coordinates, and then acquire the object pose using a pose fitting algorithm. In this paper, we conduct a comprehensive analysis of this paradigm and introduce two crucial essentials: 1) shape-sensitive and pose-invariant feature extraction for accurate correspondence prediction, and 2) outlier correspondence removal for robust pose fitting. Based on these insights, we propose a simple yet effective correspondence-based method called SpotPose, which includes two stages. During the correspondence prediction stage, pose-invariant geometric structure of objects is thoroughly exploited to facilitate shape-sensitive holistic interaction among keypoint-wise features. During the pose fitting stage, outlier scores of correspondences are explicitly predicted to facilitate efficient identification and removal of outliers. Experimental results on CAMERA25, REAL275 and HouseCat6D benchmarks demonstrate that the proposed SpotPose outperforms state-of-the-art approaches by a large margin.

****

Curriculum Direct Preference Optimization for Diffusion and Consistency Models

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, Nicu Sebe, Mubarak Shah; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2824-2834

Direct Preference Optimization (DPO) has been proposed as an effective and efficient alternative to reinforcement learning from human feedback (RLHF). In this paper, we propose a novel and enhanced version of DPO based on curriculum learning for text-to-image generation. Our method is divided into two training stages. First, a ranking of the examples generated for each prompt is obtained by employing a reward model. Then, increasingly difficult pairs of examples are sampled and provided to a text-to-image generative (diffusion or consistency) model. Generated samples that are far apart in the ranking are considered to form easy pairs, while those that are close in the ranking form hard pairs. In other words, we use the rank difference between samples as a measure of difficulty. The sampled pairs are split into batches according to their difficulty levels, which are gradually used to train the generative model. Our approach, Curriculum DPO, is compared against state-of-the-art fine-tuning approaches on nine benchmarks, outperforming the competing methods in terms of text alignment, aesthetics and human preference. Our code is available at https://github.com/CroitoruAlin/Curriculum-DPO.

****

IncEventGS: Pose-Free Gaussian Splatting from a Single Event Camera

Jian Huang, Chengrui Dong, Xuanhua Chen, Peidong Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26933-26942

Implicit neural representation and explicit 3D Gaussian Splatting (3D-GS) for novel view synthesis have achieved remarkable progress with frame-based camera (e.g. RGB and RGB-D cameras) recently. Compared to frame-based camera, a novel type

of bio-inspired visual sensor, i.e. event camera, has demonstrated advantages in high temporal resolution, high dynamic range, low power consumption and low latency, which make it being favored for many robotic applications. In this work, we present IncEventGS, an incremental 3D Gaussian Splatting reconstruction algorithm with a single event camera, without the assumption of known camera poses. To recover the 3D scene representation incrementally, we exploit the tracking and mapping paradigm of conventional SLAM pipelines for IncEventGS. Given the incoming event stream, the tracker first estimates an initial camera motion based on prior reconstructed 3D-GS scene representation. The mapper then jointly refines both the 3D scene representation and camera motion based on the previously estimated motion trajectory from the tracker. The experimental results demonstrate that IncEventGS delivers superior performance compared to prior NeRF-based methods and other related baselines, even we do not have the ground-truth camera poses. Furthermore, our method can also deliver better performance compared to state-of-the-art event visual odometry methods in terms of camera motion estimation.
****

OpenMIBOOD: Open Medical Imaging Benchmarks for Out-Of-Distribution Detection
Max Gutbrod, David Rauber, Danilo Weber Nunes, Christoph Palm; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25874-25886
The growing reliance on Artificial Intelligence (AI) in critical domains such as healthcare demands robust mechanisms to ensure the trustworthiness of these systems, especially when faced with unexpected or anomalous inputs. This paper introduces the Open Medical Imaging Benchmarks for Out-Of-Distribution Detection (OpenMIBOOD), a comprehensive framework for evaluating out-of-distribution (OOD) detection methods specifically in medical imaging contexts. OpenMIBOOD includes three benchmarks from diverse medical domains, encompassing 14 datasets divided into covariate-shifted in-distribution, near-OOD, and far-OOD categories. We evaluate 24 post-hoc methods across these benchmarks, providing a standardized reference to advance the development and fair comparison of OOD detection methods. Results reveal that findings from broad-scale OOD benchmarks in natural image domains do not translate to medical applications, underscoring the critical need for such benchmarks in the medical field. By mitigating the risk of exposing AI models to inputs outside their training distribution, OpenMIBOOD aims to support the advancement of reliable and trustworthy AI systems in healthcare. The repository is available at https://github.com/remic-othr/OpenMIBOOD.
****

Detecting Open World Objects via Partial Attribute Assignment
Muli Yang, Gabriel James Goenawan, Huaiyuan Qin, Kai Han, Xi Peng, Yanhua Yang, Hongyuan Zhu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20318-20328
Despite being trained on massive data, today's vision foundation models still fall short in detecting open world objects. Apart from recognizing known objects from training, a successful Open World Object Detection (OWOD) system must also be able to detect unknown objects never seen before, without confusing them with the backgrounds. Unlike prevailing prior works that rely on probability models to learn "objectness", we focus on learning fine-grained, class-agnostic attributes, allowing the detection of both known and unknown objects in an explainable manner. In this paper, we propose Partial Attribute Assignment (PASS), aiming to automatically select and optimize a small, relevant subset of attributes from a large attribute pool. Specifically, we model attribute selection as a Partial Optimal Transport (POT) problem between known visual objects and the attribute pool, in which more relevant attributes signify more transported mass. PASS follows a curriculum schedule that progressively selects and optimizes a targeted subset of attributes during training, promoting stability and accuracy. Our method enjoys end-to-end optimization by minimizing the POT distance and the classification loss on known visual objects, demonstrating high training efficiency and superior OWOD performance among extensive experimental evaluations.
****

FactCheXcker: Mitigating Measurement Hallucinations in Chest X-ray Report Genera

tion Models
Alice Heiman, Xiaoman Zhang, Emma Chen, Sung Eun Kim, Pranav Rajpurkar; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30787-30796
Medical vision-language models often struggle with generating accurate quantitative measurements in radiology reports, leading to hallucinations that undermine clinical reliability. We introduce FactCheXcker, a modular framework that de-hallucinates radiology report measurements by leveraging an improved query-code-update paradigm. Specifically, FactCheXcker employs specialized modules and the code generation capabilities of large language models to solve measurement queries generated based on the original report. After extracting measurable findings, the results are incorporated into an updated report. We evaluate FactCheXcker on endotracheal tube placement, which accounts for an average of 78% of report measurements, using the MIMIC-CXR dataset and 11 medical report-generation models. Our results show that FactCheXcker significantly reduces hallucinations, improves measurement precision, and maintains the quality of the original reports. Specifically, FactCheXcker improves the performance of all 11 models and achieves an average improvement of 135.0% in reducing measurement hallucinations measured by mean absolute error. Code is available at https://github.com/rajpurkarlab/FactCheXcker.
****
Neural Inverse Rendering from Propagating Light
Anagh Malik, Benjamin Attal, Andrew Xie, Matthew O'Toole, David B. Lindell; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10534-10544
We present the first system for physically based, neural inverse rendering from multi-viewpoint videos of propagating light. Our approach relies on a time-resolved extension of neural radiance caching -- a technique that accelerates inverse rendering by storing infinite-bounce radiance arriving at any point from any direction. The resulting model accurately accounts for direct and indirect light transport effects and, when applied to captured measurements from a flash lidar system, enables state-of-the-art 3D reconstruction in the presence of strong indirect light. Further, we demonstrate view synthesis of propagating light, automatic decomposition of captured measurements into direct and indirect components, as well as novel capabilities such as multi-view time-resolved relighting of captured scenes.
****
When the Future Becomes the Past: Taming Temporal Correspondence for Self-supervised Video Representation Learning
Yang Liu, Qianqian Xu, Peisong Wen, Siran Dai, Qingming Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24033-24044
The past decade has witnessed notable achievements in self-supervised learning for video tasks. Recent efforts typically adopt the Masked Video Modeling (MVM) paradigm, leading to significant progress on multiple video tasks. However, two critical challenges remain: 1) Without human annotations, the random temporal sampling introduces uncertainty, increasing the difficulty of model training. 2) Previous MVM methods primarily recover the masked patches in the pixel space, leading to insufficient information compression for downstream tasks. To address these challenges jointly, we propose a self-supervised framework that leverages Temporal Correspondence for video Representation learning (T-CoRe). For challenge 1), we propose a sandwich sampling strategy that selects two auxiliary frames to reduce reconstruction uncertainty in a two-side-squeezing manner. Addressing challenge 2), we introduce an auxiliary branch into a self-distillation architecture to restore representations in the latent space, generating high-level semantic representations enriched with temporal information. Experiments of T-CoRe consistently present superior performance across several downstream tasks, demonstrating its effectiveness for video representation learning. The code is available at https://github.com/yafeng19/T-CORE.
****

Personalized Preference Fine-tuning of Diffusion Models
Meihua Dang, Anikait Singh, Linqi Zhou, Stefano Ermon, Jiaming Song; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8020-8030
RLHF techniques like DPO can significantly improve the generation quality of text-to-image diffusion models. However, these methods optimize for a single reward that aligns model generation with population-level preferences, neglecting the nuances of individual users' beliefs or values. This lack of personalization limits the efficacy of these models. To bridge this gap, we introduce PPD, a multi-reward optimization objective that aligns diffusion models with personalized preferences. With PPD, a diffusion model learns the individual preferences of a population of users in a few-shot way, enabling generalization to unseen users. Specifically, our approach (1) leverages a vision-language model (VLM) to extract personal preference embeddings from a small set of pairwise preference examples, and then (2) incorporates the embeddings into diffusion models through cross attention. Conditioning on user embeddings, the text-to-image models are fine-tuned with the DPO objective, simultaneously optimizing for alignment with the preferences of multiple users. Empirical results demonstrate that our method effectively optimizes for multiple reward functions and can interpolate between them during inference. In real-world user scenarios, with as few as four preference examples from a new user, our approach achieves an average win rate of 76% over Stable Cascade, generating images that more accurately reflect specific user preferences.
****
DecoupledGaussian: Object-Scene Decoupling for Physics-Based Interaction
Miaowei Wang, Yibo Zhang, Weiwei Xu, Rui Ma, Changqing Zou, Daniel Morris; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11361-11372
We present DecoupledGaussian, a novel system that decouples static objects from their contacted surfaces captured in-the-wild videos, a key prerequisite for realistic Newtonian-based physical simulations. Unlike prior methods focused on synthetic data or elastic jittering along the contact surface, which prevent objects from fully detaching or moving independently, DecoupledGaussian allows for significant positional changes without being constrained by the initial contacted surface. Recognizing the limitations of current 2D inpainting tools for restoring 3D locations, our approach uses joint Poisson fields to repair and expand the Gaussians of both objects and contacted scenes after separation. This is complemented by a multi-carve strategy to refine the object's geometry. Our system enables realistic simulations of decoupling motions, collisions, and fractures driven by user-specified impulses, supporting complex interactions within and across multiple scenes. We validate DecoupledGaussian through a comprehensive user study and quantitative benchmarks. This system enhances digital interaction with objects and scenes in real-world environments, benefiting industries such as VR, robotics, and autonomous driving. Our project page is at: https://wangmiaowei.github.io/DecoupledGaussian.github.io/.
****
UniPose: A Unified Multimodal Framework for Human Pose Comprehension, Generation and Editing
Yiheng Li, Ruibing Hou, Hong Chang, Shiguang Shan, Xilin Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27805-27815
Human pose plays a crucial role in the digital age. While recent works have achieved impressive progress in understanding and generating human poses, they often support only a single modality of control signals and operate in isolation, limiting their application in real-world scenarios. This paper presents UniPose, a framework employing Large Language Models (LLMs) to comprehend, generate, and edit human poses across various modalities, including images, text, and 3D SMPL poses. Specifically, we apply a pose tokenizer to convert 3D poses into discrete pose tokens, enabling seamless integration into the LLM within a unified vocabulary. To further enhance the fine-grained pose perception capabilities, we facili

tate UniPose with a mixture of visual encoders, among them a pose-specific visua
l encoder. Benefiting from a unified learning strategy, UniPose effectively tran
sfers knowledge across different pose-relevant tasks, adapts to unseen tasks, an
d exhibits extended capabilities. This work serves as the first attempt at build
ing a general-purpose framework for pose comprehension, generation, and editing.
 Extensive experiments highlight UniPose's competitive and even superior perform
ance across various pose-relevant tasks. The code is available at https://github
.com/liyiheng23/UniPose.
****

POMP: Physics-consistent Motion Generative Model through Phase Manifolds
Bin Ji, Ye Pan, Zhimeng Liu, Shuai Tan, Xiaogang Jin, Xiaokang Yang; Proceedings
 of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 226
90-22701
Numerous researches on real-time motion generation primarily focus on kinematic
aspects, often resulting in physically implausible outcomes. In this paper, we p
resent POMP ("\underline P hysics-c\underline O nsistent Human \underline M otio
n \underline P rior through Phase Manifolds"), a novel kinematics-based framewor
k that synthesizes physically consistent motions by leveraging phase manifolds t
o align motion priors with physics constraints. POMP operates as a frame-by-fram
e autoregressive model with three core components: a diffusion-based kinematic m
odule, a simulation-based dynamic module, and a phase encoding module. At each t
imestep, the kinematic module generates an initial target pose, which is subsequ
ently refined by the dynamic module to simulate human-environment interactions.
Although the physical simulation ensures adherence to physical laws, it may comp
romise the kinematic rationality of the posture. Consequently, directly using th
e simulated result for subsequent frame prediction may lead to cumulative errors
. To address this, the phase encoding module performs semantic alignment in the
phase manifold. Moreover, we present a pipeline in Unity for generating terrain
maps and capturing full-body motion impulses from existing motion capture (MoCap
) data. The collected terrain topology and motion impulse data facilitate the tr
aining of POMP, enabling it to robustly respond to underlying contactforces and
applied dynamics. Extensive evaluations demonstrate the efficacy of POMP across
various contexts, terrains, and physical interactions.
****

NN-Former: Rethinking Graph Structure in Neural Architecture Representation
Ruihan Xu, Haokui Zhang, Yaowei Wang, Wei Zeng, Shiliang Zhang; Proceedings of t
he Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10004-10
014
The growing use of deep learning necessitates efficient network design and deplo
yment, making neural predictors vital for estimating attributes such as accuracy
 and latency. Recently, Graph Neural Networks (GNNs) and transformers have shown
 promising performance in representing neural architectures. However, each metho
d has its disadvantages. GNNs lack the capabilities to represent complicated fea
tures, while transformers face poor generalization when the depth of architectur
e grows. To mitigate the above problems, we rethink neural architecture topology
 and show that sibling nodes are pivotal while overlooked in previous research.
Thus we propose a novel predictor leveraging the strengths of GNNs and transform
ers to learn the enhanced topology. We introduce a novel token mixer that consid
ers siblings, and a new channel mixer named bidirectional graph isomorphism feed
-forward network. Our approach consistently achieves promising performance in bo
th accuracy and latency prediction, providing valuable insights for learning Dir
ected Acyclic Graph (DAG) topology. The code is available at https://github.com/
XuRuihan/NNFormer.
****

DashGaussian: Optimizing 3D Gaussian Splatting in 200 Seconds
Youyu Chen, Junjun Jiang, Kui Jiang, Xiao Tang, Zhihao Li, Xianming Liu, Yinyu N
ie; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)
, 2025, pp. 11146-11155
3D Gaussian Splatting (3DGS) renders pixels by rasterizing Gaussian primitives,
where the rendering resolution and the primitive number, concluded as the optimi

zation complexity, dominate the time cost in primitive optimization. In this pap
er, we propose DashGaussian, a scheduling scheme over the optimization complexit
y of 3DGS that strips redundant complexity to accelerate 3DGS optimization. Spec
ifically, we formulate 3DGS optimization as progressively fitting 3DGS to higher
 levels of frequency components in the training views, and propose a dynamic ren
dering resolution scheme that largely reduces the optimization complexity based
on this formulation. Besides, we argue that a specific rendering resolution shou
ld cooperate with a proper primitive number for a better balance between computi
ng redundancy and fitting quality, where we schedule the growth of the primitive
s to synchronize with the rendering resolution. Extensive experiments show that
our method accelerates the optimization of various 3DGS backbones by 45.7% on av
erage while preserving the rendering quality.
****

Reasoning to Attend: Try to Understand How <SEG> Token Works
Rui Qian, Xin Yin, Dejing Dou; Proceedings of the Computer Vision and Pattern Re
cognition Conference (CVPR), 2025, pp. 24722-24731
Current Large Multimodal Models (LMMs) empowered visual grounding typically rely
 on \texttt <SEG> tokens as a text prompt to jointly optimize the vision-langua
ge model (e.g., LLaVA) and the downstream task-specific model (e.g., SAM). Howev
er, we observe that little research has looked into how it works. In this work,
we first visualize the similarity maps, which are obtained by computing the sema
ntic similarity between the \texttt <SEG> token and the image token embeddings
derived from the last hidden layer in both the LLaVA encoder and SAM decoder. In
triguingly, we have found that a striking consistency holds in terms of activati
on responses in the similarity map, which reveals that what the \texttt <SEG> t
oken contributes to is semantic similarity within image-text pairs. Specificall
y, the \texttt <SEG> token, a placeholder expanded in text vocabulary, extensiv
ely queries among individual tokenized image patches to match the semantics of a
n object from text to the paired image, while the Large Language Models (LLMs) a
re being fine-tuned. Upon the above findings, we present \toolname, which facili
tates LMMs' resilient REAsoning capability of where to attenD under the guidance
 of highly activated points borrowed from similarity maps. Remarkably, READ feat
ures an intuitive design, Similarity as Points module (SasP), which can be seaml
essly applied to \texttt <SEG> -like paradigms in a plug-and-play fashion. Also,
 extensive experiments have been conducted on ReasonSeg and RefCOCO(+/g) dataset
s. To validate whether READ suffers from catastrophic forgetting of previous ski
lls after fine-tuning, we further assess its generation ability on an augmented
FP-RefCOCO(+/g) dataset. All codes and models are publicly available at \href ht
tps://github.com/rui-qian/READ https://github.com/rui-qian/READ .
****

ReSpec: Relevance and Specificity Grounded Online Filtering for Learning on Vide
o-Text Data Streams
Chris Dongjoo Kim, Jihwan Moon, Sangwoo Moon, Heeseung Yun, Sihaeng Lee, Anirudd
ha Kembhavi, Soonyoung Lee, Gunhee Kim, Sangho Lee, Christopher Clark; Proceedin
gs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2
9040-29049
The rapid growth of video-text data presents challenges in storage and computati
on during training. Online learning, which processes streaming data in real-time
, offers a promising solution to these issues while also allowing swift adaptati
ons in scenarios demanding real-time responsiveness. One strategy to enhance the
 efficiency and effectiveness of learning involves identifying and prioritizing
data that enhances performance on target downstream tasks. We propose Relevance
and Specificity-based online filtering framework (ReSpec) that selects data base
d on four criteria: (i) modality alignment for clean data, (ii) task relevance f
or target focused data, (iii) specificity for informative and detailed data, and
 (iv) efficiency for low-latency processing. Relevance is determined by the prob
abilistic alignment of incoming data with downstream tasks, while specificity em
ploys the distance to a root embedding representing the least specific data as a
n efficient proxy for informativeness. By establishing reference points from tar
get task data, ReSpec filters incoming data in real-time, eliminating the need f

or extensive storage and compute. Evaluating on large-scale datasets WebVid2M and VideoCC3M, ReSpec attains state-of-the-art performance on five zero-shot video retrieval tasks, using as little as 5% of the data while incurring minimal compute. The source code is available at https://github.com/cdjkim/ReSpec.
****

A Unified Image-Dense Annotation Generation Model for Underwater Scenes
Hongkai Lin, Dingkang Liang, Zhenghao Qi, Xiang Bai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 961-970
Underwater dense prediction, especially depth estimation and semantic segmentation, is crucial for gaining a comprehensive understanding of underwater scenes. Nevertheless, high-quality and large-scale underwater datasets with dense annotations remain scarce because of the complex environment and the exorbitant data collection costs. This paper proposes a unified Text-to-Image and DEnse annotation generation method (TIDE) for underwater scenes. It relies solely on text as input to simultaneously generate realistic underwater images and multiple highly consistent dense annotations. Specifically, we unify the generation of text-to-image and text-to-dense annotations within a single model. The Implicit Layout Sharing mechanism (ILS) and cross-modal interaction method called Time Adaptive Normalization (TAN) are introduced to jointly optimize the consistency between image and dense annotations. We synthesize a large-scale underwater dataset using TIDE to validate the effectiveness of our method in underwater dense prediction tasks. The results demonstrate that our method effectively improves the performance of existing underwater dense prediction models and mitigates the scarcity of underwater data with dense annotations. We hope our method can offer new perspectives on alleviating data scarcity issues in other fields. The code is available at https://github.com/HongkLin/TIDE.
****

PACT: Pruning and Clustering-Based Token Reduction for Faster Visual Language Models
Mohamed Dhouib, Davide Buscaldi, Sonia Vanier, Aymen Shabou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14582-14592
Visual Language Models require substantial computational resources for inference due to the additional input tokens needed to represent visual information. However, these visual tokens often contain redundant and unimportant information, resulting in an unnecessarily high number of tokens. To address this, we introduce PACT, a method that reduces inference time and memory usage by pruning irrelevant tokens and merging visually redundant ones at an early layer of the language model. Our approach uses a novel importance metric to identify unimportant tokens without relying on attention scores, making it compatible with FlashAttention. We also propose a novel clustering algorithm, called Distance Bounded Density Peak Clustering, which efficiently clusters visual tokens while constraining the distances between elements within a cluster by a predefined threshold. We demonstrate the effectiveness of PACT through extensive experiments.
****

R-SCoRe: Revisiting Scene Coordinate Regression for Robust Large-Scale Visual Localization
Xudong Jiang, Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Marc Pollefeys; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11536-11546
Learning-based visual localization methods that use scene coordinate regression (SCR) offer the advantage of smaller map sizes. However, on datasets with complex illumination changes or image-level ambiguities, it remains a less robust alternative to feature matching methods. This work aims to close the gap. We introduce a covisibility graph-based global encoding learning and data augmentation strategy, along with a depth-adjusted reprojection loss to facilitate implicit triangulation. Additionally, we revisit the network architecture and local feature extraction module. Our method achieves state-of-the-art on challenging large-scale datasets without relying on network ensembles or 3D supervision. On Aachen Day-Night, we are 10x more accurate than previous SCR methods with similar map sizes and require at least 5x smaller map sizes than any other SCR method while stil

l delivering superior accuracy. Code is available at: https://github.com/cvg/scr
studio.
****

DynRefer: Delving into Region-level Multimodal Tasks via Dynamic Resolution
Yuzhong Zhao, Feng Liu, Yue Liu, Mingxiang Liao, Chen Gong, Qixiang Ye, Fang Wan
; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR),
2025, pp. 24742-24752
One important task of multimodal models is to translate referred image regions t
o human preferred language descriptions. Existing methods, however, ignore the r
esolution adaptability needs of different tasks, which hinders them to find out
precise language descriptions. In this study, we propose a DynRefer approach, to
 pursue high-accuracy region-level referring through mimicking the resolution ad
aptability of human visual cognition. During training, DynRefer stochastically a
ligns language descriptions of multimodal tasks with images of multiple resoluti
ons, which are constructed by nesting a set of random views around the referred
region. This process essentially constructs a set of region representations, whe
re suitable representations for specific tasks can be matched. During inference,
 DynRefer performs selectively multimodal referring by sampling proper region re
presentations for tasks from the set of views based on image and task priors. Th
is allows the visual information for referring to better match human preferences
, thereby improving the representational adaptability of region-level multimodal
 models. Experiments show that DynRefer brings mutual improvement upon broad tas
ks including region-level captioning, open-vocabulary region recognition and att
ribute detection. Furthermore, DynRefer achieves state-of-the-art results on mul
tiple region-level multimodal tasks using a single model. Code is available at h
ttps://github.com/callsys/DynRefer.
****

Playing the Fool: Jailbreaking LLMs and Multimodal LLMs with Out-of-Distribution
 Strategy
Joonhyun Jeong, Seyun Bae, Yeonsung Jung, Jaeryong Hwang, Eunho Yang; Proceeding
s of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29
937-29946
Despite the remarkable versatility of Large Language Models (LLMs) and Multimoda
l LLMs (MLLMs) to generalize across both language and vision tasks, LLMs and MLL
Ms have shown vulnerability to jailbreaking, generating textual outputs that und
ermine safety, ethical, and bias standards when exposed to harmful or sensitive
inputs. With the recent advancement of safety alignment via preference-tuning fr
om human feedback, LLMs and MLLMs have been equipped with safety guardrails to y
ield safe, ethical, and fair responses with regard to harmful inputs. However, d
espite the significance of safety alignment, research on the vulnerabilities rem
ains largely underexplored. In this paper, we investigate the unexplored vulnera
bility of the safety alignment, examining its ability to consistently provide sa
fety guarantees for out-of-distribution(OOD)-ifying harmful inputs that may fall
 outside the aligned data distribution. Our key observation is that OOD-ifying t
he vanilla harmful inputs highly increases the uncertainty of the model to disce
rn the malicious intent within the input, leading to a higher chance of being ja
ilbroken. Exploiting this vulnerability, we propose JOOD, a new Jailbreak framew
ork via OOD-ifying inputs beyond the safety alignment. We explore various off-th
e-shelf visual and textual transformation techniques for OOD-ifying the harmful
inputs. Notably, we observe that even simple mixing-based techniques such as ima
ge mixup prove highly effective in increasing the uncertainty of the model, ther
eby facilitating the bypass of the safety alignment. Experiments across diverse
jailbreak scenarios demonstrate that JOOD effectively jailbreaks recent propriet
ary LLMs and MLLMs such as GPT-4 and o1 with high attack success rate, which pre
vious attack approaches have consistently struggled to jailbreak. Code is availa
ble at https://github.com/naver-ai/JOOD.
****

Style Evolving along Chain-of-Thought for Unknown-Domain Object Detection
Zihao Zhang, Aming Wu, Yahong Han; Proceedings of the Computer Vision and Patter
n Recognition Conference (CVPR), 2025, pp. 14225-14234

Recently, a task of Single-Domain Generalized Object Detection (Single-DGOD) is proposed, aiming to generalize a detector to multiple unknown domains never seen before during training. Due to the unavailability of target-domain data, some methods leverage the multimodal capabilities of vision-language models, using textual prompts to estimate cross-domain information, enhancing the model's generalization capability. These methods typically use a single textual prompt, referred to as the one-step prompt method. However, when dealing with complex styles, such as the combination of rain and night, we observe that the performance of the one-step prompt method tends to be relatively weak. The reason may be that many scenes incorporate a single style and a combination of multiple styles. The one-step prompt method may not effectively synthesize combined information involving various styles. To address this limitation, we propose a new method, i.e., Style Evolving along Chain-of-Thought, which aims to progressively integrate and expand style information along the chain of thought, enabling the continual evolution of styles. Specifically, by progressively refining style descriptions and guiding the diverse evolution of styles, this method enhances the simulation of various style characteristics, enabling the model to learn and adapt to subtle differences more effectively. Additionally, it exposes the model to a broader range of style features with different data distributions, thereby enhancing its generalization capability in unseen domains. The significant performance gains over five adverse-weather scenarios and the Real to Art benchmark demonstrate the superiorities of our method. Our code is available at https://github.com/ZZ2490/SE-COT.

****

NTR-Gaussian: Nighttime Dynamic Thermal Reconstruction with 4D Gaussian Splatting Based on Thermodynamics

Kun Yang, Yuxiang Liu, Zeyu Cui, Yu Liu, Maojun Zhang, Shen Yan, Qing Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 691-700

Thermal infrared imaging enables a non-invasive measurement of the surface temperature of objects with all-weather applicability. Leveraging such techniques for 3D reconstruction can accurately reflect the temperature distribution of a scene, thereby supporting applications such as building monitoring and energy management. However, existing approaches predominantly focus on static 3D reconstruction for a single time period, overlooking the dynamic nature of thermal radiation phenomena, and failing to predict or analyze temperature variations over time. In this paper, we introduce a novel method, termed NTR-Gaussian, grounded in thermodynamics to address the challenge of nighttime dynamic thermal reconstruction using 4D Gaussian Splatting. Specifically, We utilize neural networks to predict thermodynamic parameters, such as emissivity, convective heat transfer coefficient, and heat capacity, etc. By means of integration, we numerically solve the infrared temperature of the scene at each moment during the night, so as to predict the temperature of the nighttime scene more accurately. To further advance research in this domain, we release a comprehensive dataset of dynamic thermal reconstruction spanning four distinct regions. Extensive experiments demonstrate that NTR-Gaussian significantly outperforms comparison methods in thermal reconstruction, achieving a predicted temperature error within 1 degree Celsius.

****

OmniSplat: Taming Feed-Forward 3D Gaussian Splatting for Omnidirectional Images with Editable Capabilities

Suyoung Lee, Jaeyoung Chung, Kihoon Kim, Jaeyoo Huh, Gunhee Lee, Minsoo Lee, Kyoung Mu Lee; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16356-16365

Feed-forward 3D Gaussian splatting (3DGS) models have gained significant popularity due to their ability to generate scenes immediately without needing per-scene optimization. Although omnidirectional images are becoming more popular since they reduce the computation required for image stitching to composite a holistic scene, existing feed-forward models are only designed for perspective images. The unique optical properties of omnidirectional images make it difficult for feature encoders to correctly understand the context of the image and make the Gaus

sian non-uniform in space, which hinders the image quality synthesized from novel views. We propose OmniSplat, a training-free fast feed-forward 3DGS generation framework for omnidirectional images. We adopt a Yin-Yang grid and decompose images based on it to reduce the domain gap between omnidirectional and perspective images. The Yin-Yang grid can use the existing CNN structure as it is, but its quasi-uniform characteristic allows the decomposed image to be similar to a perspective image, so it can exploit the strong prior knowledge of the learned feed-forward network. OmniSplat demonstrates higher reconstruction accuracy than existing feed-forward networks trained on perspective images. The code is available on: https://github.com/esw0116/OmniSplat.
****

VideoWorld: Exploring Knowledge Learning from Unlabeled Videos
Zhongwei Ren, Yunchao Wei, Xun Guo, Yao Zhao, Bingyi Kang, Jiashi Feng, Xiaojie Jin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29029-29039
This work explores whether a deep generative model can learn complex knowledge solely from visual input, in contrast to the prevalent focus on text-based models like large language models (LLMs). We develop VideoWorld, an auto-regressive video generation model trained on unlabeled video data, and test its knowledge acquisition abilities in video-based Go and robotic control tasks. Our experiments reveal two key findings: (1) video-only training provides sufficient information for learning knowledge, including rules, reasoning and planning capabilities, and (2) the representation of visual change is crucial for knowledge acquisition. To improve both the efficiency and efficacy of this process, we introduce the Latent Dynamics Model (LDM) as a key component of VideoWorld. Remarkably, VideoWorld reaches a 5-dan professional level in the Video-GoBench with just a 300-million-parameter model, without relying on search algorithms or reward mechanisms typical in reinforcement learning. In robotic tasks, VideoWorld effectively learns diverse control operations and generalizes across environments, approaching the performance of oracle models in CALVIN and RLBench. This study opens new avenues for knowledge acquisition from visual data, with all code, data, and models open-sourced for further research.
****

FSHNet: Fully Sparse Hybrid Network for 3D Object Detection
Shuai Liu, Mingyue Cui, Boyang Li, Quanmin Liang, Tinghe Hong, Kai Huang, Yunxiao Shan, Kai Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8900-8909
Fully sparse 3D detectors have recently gained significant attention due to their efficiency in long-range detection. However, sparse 3D detectors extract features only from non-empty voxels, which impairs long-range interactions and causes the center feature missing. The former weakens the feature extraction capability, while the latter hinders network optimization. To address these challenges, we introduce the Fully Sparse Hybrid Network (FSHNet). FSHNet incorporates a proposed SlotFormer block to enhance the long-range feature extraction capability of existing sparse encoders. The SlotFormer divides sparse voxels using a slot partition approach, which, compared to traditional window partition, provides a larger receptive field. Additionally, we propose a dynamic sparse label assignment strategy to deeply optimize the network by providing more high-quality positive samples. To further enhance performance, we introduce a sparse upsampling module to refine downsampled voxels, preserving fine-grained details crucial for detecting small objects. Extensive experiments on the Waymo, nuScenes, and Argoverse2 benchmarks demonstrate the effectiveness of FSHNet. The code will be publicly available.
****

3D-SLNR: A Super Lightweight Neural Representation for Large-scale 3D Mapping
Chenhui Shi, Fulin Tang, Ning An, Yihong Wu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27233-27242
We propose 3D-SLNR, a new and ultra-lightweight neural representation with outstanding performance for large-scale 3D mapping. The representation defines a global signed distance function (SDF) in near-surface space based on a set of band-l

imited local SDFs anchored at support points sampled from point clouds. These SD
Fs are parameterized only by a tiny multi-layer perceptron (MLP) with no latent
features, and the state of each SDF is modulated by three learnable geometric pr
operties: position, rotation, and scaling, which make the representation adapt t
o complex geometries. Then, we develop a novel parallel algorithm tailored for t
his unordered representation to efficiently detect local SDFs where each sampled
 point is located, allowing for real-time updates of local SDF states during tra
ining. Additionally, a prune-and-expand strategy is introduced to enhance adapta
bility further. The synergy of our low-parameter model and its adaptive capabili
ties results in an extremely compact representation with excellent expressivenes
s. Extensive experiments demonstrate that our method achieves state-of-the-art r
econstruction performance with less than 1/5 of the memory footprint compared wi
th previous advanced methods.
****

UniVAD: A Training-free Unified Model for Few-shot Visual Anomaly Detection
Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, Jinqiao Wang; Proc
eedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025,
pp. 15194-15203
Visual Anomaly Detection (VAD) aims to identify abnormal samples in images that
deviate from normal patterns, covering multiple domains, including industrial, l
ogical, and medical fields. Due to the domain gaps between these fields, existin
g VAD methods are typically tailored to each domain, with specialized detection
techniques and model architectures that are difficult to generalize across diffe
rent domains. Moreover, even within the same domain, current VAD approaches ofte
n require large amounts of normal samples to train class-specific models, result
ing in poor generalizability and hindering unified evaluation across domains. To
 address this issue, we propose a generalized few-shot VAD method, UniVAD, capab
le of detecting anomalies across various domains, with a training-free unified m
odel. UniVAD only needs few normal samples as references during testing to detec
t anomalies in previously unseen objects, without training on the specific domai
n. Specifically, UniVAD employs a Contextual Component Clustering (C3) module ba
sed on clustering and vision foundation models to segment components within the
image accurately, and leverages Component-Aware Patch Matching (CAPM) and Graph-
Enhanced Component Modeling (GECM) modules to detect anomalies at different sema
ntic levels, which are aggregated to produce the final detection result. We cond
uct experiments on nine datasets spanning industrial, logical, and medical field
s, and the results demonstrate that UniVAD achieves state-of-the-art performance
 in few-shot anomaly detection tasks across multiple domains, outperforming doma
in-specific anomaly detection models. Code is available at https://github.com/Fa
ntasticGNU/UniVAD.
****

STINR: Deciphering Spatial Transcriptomics via Implicit Neural Representation
Yisi Luo, Xile Zhao, Kai Ye, Deyu Meng; Proceedings of the Computer Vision and P
attern Recognition Conference (CVPR), 2025, pp. 25930-25939
Spatial transcriptomics (ST) are emerging technologies that reveal spatial distr
ibutions of gene expressions within tissues, serving as important ways to uncove
r biological insights. However, the irregular spatial profiles and variability o
f genes make it challenging to integrate spatial information with gene expressio
n under a computational framework. Current algorithms mostly utilize spatial gra
ph neural networks to encode spatial information, which may incur increased comp
utational costs and may not be flexible enough to depict complex spatial configu
rations. In this study, we introduce a concise yet effective representation fram
ework, STINR, for deciphering ST data. STINR leverages an implicit neural repres
entation (INR) to continuously represent ST data, which efficiently characterize
s spatial and slice-wise correlations of ST data by inheriting the implicit smoo
thness of INR. STINR allows easier integration of multiple slices and multi-omic
s without any alignment, and serves as a potent tool for various biological task
s including gene imputation, gene denoising, spatial domain detection, and cell-
type deconvolution stemed from ST data. In particular, STINR identifies the thin
nest cortex layer in the dorsolateral prefrontal cortex which previous methods w

ere unable to achieve, and more accurately identifies tumor regions in the human squamous cell carcinoma, showcasing its practical value for biological discoveries.

****

Remote Photoplethysmography in Real-World and Extreme Lighting Scenarios

Hang Shao, Lei Luo, Jianjun Qian, Mengkai Yan, Shuo Chen, Jian Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10858-10867

Physiological activities can be manifested by the sensitive changes in facial imaging. While they are barely observable to our eyes, computer vision manners can, and the derived remote photoplethysmography (rPPG) has shown considerable promise. However, existing studies mainly rely on spatial skin recognition and temporal rhythmic interactions, so they focus on identifying explicit features under ideal light conditions, but perform poorly in-the-wild with intricate obstacles and extreme illumination exposure. In this paper, we propose an end-to-end video transformer model for rPPG. It strives to eliminate complex and unknown external time-varying interferences, whether they are sufficient to occupy subtle biosignal amplitudes or exist as periodic perturbations that hinder network training. In the specific implementation, we utilize global interference sharing, subject background reference, and self-supervised disentanglement to eliminate interference, and further guide learning based on spatiotemporal filtering, reconstruction guidance, and frequency domain and biological prior constraints to achieve effective rPPG. To the best of our knowledge, this is the first robust rPPG model for real outdoor scenarios based on natural face videos, and is lightweight to deploy. Extensive experiments show the competitiveness and performance of our model in rPPG prediction across datasets and scenes.

****

Multi-Modal Contrastive Masked Autoencoders: A Two-Stage Progressive Pre-training Approach for RGBD Datasets

Muhammad Abdullah Jamal, Omid Mohareri; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17947-17957

In this paper, we propose a new progressive pre-training method for image understanding tasks which leverages RGB-D datasets. The method utilizes Multi-Modal Contrastive Masked Autoencoder and Denoising techniques. Our proposed approach consists of two stages. In the first stage, we pre-train the model using contrastive learning to learn cross-modal representations. In the second stage, we further pre-train the model using masked autoencoding and denoising/noise prediction used in diffusion models. Masked autoencoding focuses on reconstructing the missing patches in the input modality using local spatial correlations, while denoising learns high frequency components of the input data. Moreover, it incorporates global distillation in the second stage by leveraging the knowledge acquired in stage one. Our approach is scalable, robust and suitable for pre-training RGB-D datasets. Extensive experiments on multiple datasets such as ScanNet, NYUv2 and SUN RGB-D show the efficacy and superior performance of our approach. Specifically, we show an improvement of +1.3% mIoU against Mask3D on ScanNet semantic segmentation. We further demonstrate the effectiveness of our approach in low-data regime by evaluating it for semantic segmentation task against the state-of-the-art methods.

****

JTD-UAV: MLLM-Enhanced Joint Tracking and Description Framework for Anti-UAV Systems

Yifan Wang, Jian Zhao, Zhaoxin Fan, Xin Zhang, Xuecheng Wu, Yudian Zhang, Lei Jin, Xinyue Li, Gang Wang, Mengxi Jia, Ping Hu, Zheng Zhu, Xuelong Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1633-1644

Unmanned Aerial Vehicles (UAVs) are widely adopted across various fields, yet they raise significant privacy and safety concerns, demanding robust monitoring solutions. Existing anti-UAV methods primarily focus on position tracking but fail to capture UAV behavior and intent. To address this, we introduce a novel task--UAV Tracking and Intent Understanding (UTIU)--which aims to track UAVs while in

ferring and describing their motion states and intent for a more comprehensive m
onitoring approach. To tackle the task, we propose JTD-UAV, the first joint trac
king, and intent description framework based on large language models. Our dual-
branch architecture integrates UAV tracking with Visual Question Answering (VQA)
, allowing simultaneous localization and behavior description. To benchmark this
 task, we introduce the TDUAV dataset, the largest dataset for joint UAV trackin
g and intent understanding, featuring 1,328 challenging video sequences, over 16
3K annotated thermal frames, and 3K VQA pairs. Our benchmark demonstrates the ef
fectiveness of JTD-UAV, and both the dataset and code will be publicly available
.
****

HaWoR: World-Space Hand Motion Reconstruction from Egocentric Videos
Jinglei Zhang, Jiankang Deng, Chao Ma, Rolandos Alexandros Potamias; Proceedings
 of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 180
5-1815
Despite the advent in 3D hand pose estimation, current methods predominantly foc
us on single-image 3D hand reconstruction in the camera frame, overlooking the w
orld-space motion of the hands. Such limitation prohibits their direct use in eg
ocentric video settings, where hands and camera are continuously in motion. In t
his work, we propose HaWoR, a high-fidelity method for hand motion reconstructio
n in world coordinates from egocentric videos. We propose to decouple the task b
y reconstructing the hand motion in the camera space and estimating the camera t
rajectory in the world coordinate system. To achieve precise camera trajectory e
stimation, we propose an adaptive egocentric SLAM framework that addresses the s
hortcomings of traditional SLAM methods, providing robust performance under chal
lenging camera dynamics. To ensure robust hand motion trajectories, even when th
e hands move out of view frustum, we devise a novel motion infiller network that
 effectively completes the missing frames of the sequence. Through extensive qua
ntitative and qualitative evaluations, we demonstrate that HaWoR achieves state-
of-the-art performance on both hand motion reconstruction and world-frame camera
 trajectory estimation under different egocentric benchmark datasets. Code and m
odels are available on \href https://hawor-project.github.io/ https://hawor-pro
ject.github.io/ .
****

Font-Agent: Enhancing Font Understanding with Large Language Models
Yingxin Lai, Cuijie Xu, Haitian Shi, Guoqing Yang, Xiaoning Li, Zhiming Luo, Sha
ozi Li; Proceedings of the Computer Vision and Pattern Recognition Conference (C
VPR), 2025, pp. 19670-19680
The rapid development of generative models has significantly advanced font gener
ation. However, limited exploration has been devoted to the evaluation and inter
pretability of graphical fonts. Existing quality assessment models can only prov
ide basic visual analyses, such as recognizing clarity and brightness, without i
n-depth explanations. To address these limitations, we first constructed a large
-scale multimodal dataset named the Diversity Font Dataset (DFD), comprising 135
,000 font-text pairs. This dataset encompasses a wide range of generated font ty
pes and annotations, including language descriptions and quality assessments, th
us providing a robust foundation for training and evaluating font analysis model
s. Based on this dataset, we developed a font agent built upon a Vision-Language
 Model (VLM) aiming to enhance font quality assessment and offer interpretable q
uestion-answering capabilities. Alongside the original visual encoder in VLM, we
 integrated an Edge-Aware Traces (EAT) module to capture detailed edge informati
on of font strokes and components. Furthermore, we introduced a Dynamic Direct P
reference Optimization (D-DPO) strategy to facilitate efficient model fine-tunin
g. Experimental results demonstrate that Font-Agent achieves state-of-the-art pe
rformance on the established dataset. To further evaluate the generalization abi
lity of our algorithm, we conducted additional experiments on several public dat
asets. The results highlight the notable advantage of Font-Agent in both assessi
ng the quality of generated fonts and comprehending their content.
****

Secret Lies in Color: Enhancing AI-Generated Images Detection with Color Distrib

ution Analysis
Zexi Jia, Chuanwei Huang, Yeshuang Zhu, Hongyan Fei, Xiaoyue Duan, Zhiqiang Yuan , Ying Deng, Jiapei Zhang, Jinchao Zhang, Jie Zhou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13445-13454
The advancement of Generative Adversarial Networks (GANs) and diffusion models s ignificantly enhances the realism of synthetic images, driving progress in image processing and creative design. However, this progress also necessitates the de velopment of effective detection methods, as synthetic images become increasingl y difficult to distinguish from real ones. This difficulty leads to societal iss ues, such as the spread of misinformation, identity theft, and online fraud. Whi le previous detection methods perform well on public benchmarks, they struggle w ith our benchmark, FakeART, particularly when dealing with the latest models and cross-domain tasks (e.g., photo-to-painting). To address this challenge, we dev elop a new synthetic image detection technique based on color distribution. Unli ke real images, synthetic images often exhibit uneven color distribution. By emp loying color quantization and restoration techniques, we analyze the color diffe rences before and after image restoration. We discover and prove that these diff erences closely relate to the uniformity of color distribution. Based on this fi nding, we extract effective color features and combine them with image features to create a detection model with only 1.4 million parameters. This model achieve s state-of-the-art results across various evaluation benchmarks, including the c hallenging FakeART dataset.
****

RADIOv2.5: Improved Baselines for Agglomerative Vision Foundation Models
Greg Heinrich, Mike Ranzinger, Hongxu Yin, Yao Lu, Jan Kautz, Andrew Tao, Bryan Catanzaro, Pavlo Molchanov; Proceedings of the Computer Vision and Pattern Recog nition Conference (CVPR), 2025, pp. 22487-22497
Agglomerative models have recently emerged as a powerful approach to training vi sion foundation models, leveraging multi-teacher distillation from existing mode ls such as CLIP, DINO, and SAM. This strategy enables the efficient creation of robust models, combining the strengths of individual teachers while significantl y reducing computational and resource demands. In this paper, we thoroughly anal yze state-of-the-art agglomerative models, identifying critical challenges inclu ding resolution mode shifts, teacher imbalance, idiosyncratic teacher artifacts, and an excessive number of output tokens. To address these issues, we propose s everal novel solutions: multi-resolution training, mosaic augmentation, and impr oved balancing of teacher loss functions. Specifically, in the context of Vision Language Models, we introduce a token compression technique to maintain high-re solution information within a fixed token count. We release our top-performing v ariants at multiple scales (-B, -L, -H, and -g), along with inference code and p retrained weights.
****

High Temporal Consistency through Semantic Similarity Propagation in Semi-Superv ised Video Semantic Segmentation for Autonomous Flight
Cédric Vincent, Taehyoung Kim, Henri Meeß; Proceedings of the Computer Vision an d Pattern Recognition Conference (CVPR), 2025, pp. 1461-1471
Semantic segmentation from RGB cameras is essential to the perception of autonom ous flying vehicles. The stability of predictions through the captured videos is paramount to their reliability and, by extension, to the trustworthiness of the agents. In this paper, we propose a lightweight video semantic segmentation app roach--suited to onboard real-time inference--achieving high temporal consistenc y on aerial data through Semantic Similarity Propagation across frames. SSP temp orally propagates the predictions of an efficient image segmentation model with global registration alignment to compensate for camera movements. It combines th e current estimation and the prior prediction with linear interpolation using we ights computed from the features similarities of the two frames. Because data av ailability is a challenge in this domain, we propose a consistency-aware Knowled ge Distillation training procedure for sparsely labeled datasets with few annota tions. Using a large image segmentation model as a teacher to train the efficien t SSP, we leverage the strong correlations between labeled and unlabeled frames

in the same training videos to obtain high-quality supervision on all frames. KD-SSP obtains a significant temporal consistency increase over the base image segmentation model of 12.5% and 6.7% TC on UAVid and RuralScapes respectively, with higher accuracy and comparable inference speed. On these aerial datasets, KD-SSP provides a superior segmentation quality and inference speed trade-off than other video methods proposed for general applications and shows considerably higher consistency. The code will be made publicly available upon acceptance.
****

Cross-Modal and Uncertainty-Aware Agglomeration for Open-Vocabulary 3D Scene Understanding
Jinlong Li, Cristiano Saltori, Fabio Poiesi, Nicu Sebe; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19390-19400
The lack of a large-scale 3D-text corpus has led recent works to distill open-vocabulary knowledge from vision-language models (VLMs). However, these methods typically rely on a single VLM to align the feature spaces of 3D models within a common language space, which limits the potential of 3D models to leverage the diverse spatial and semantic capabilities encapsulated in various foundation models. In this paper, we propose Cross-modal and Uncertainty-aware Agglomeration for Open-vocabulary 3D Scene Understanding dubbed CUA-O3D, the first model to integrate multiple foundation models--such as CLIP, DINOv2, and Stable Diffusion--into 3D scene understanding. We further introduce a deterministic uncertainty estimation to adaptively distill and harmonize the heterogeneous 2D feature embeddings from these models. Our method addresses two key challenges: (1) incorporating semantic priors from VLMs alongside the geometric knowledge of spatially-aware vision foundation models, and (2) using a novel deterministic uncertainty estimation to capture model-specific uncertainties across diverse semantic and geometric sensitivities, helping to reconcile heterogeneous representations during training. Extensive experiments on ScanNetV2 and Matterport3D demonstrate that our method not only advances open-vocabulary segmentation but also achieves robust cross-domain alignment and competitive spatial perception capabilities.
****

Generative Gaussian Splatting for Unbounded 3D City Generation
Haozhe Xie, Zhaoxi Chen, Fangzhou Hong, Ziwei Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6111-6120
3D city generation with NeRF-based methods shows promising generation results but is computationally inefficient. Recently 3D Gaussian splatting (3D-GS) has emerged as a highly efficient alternative for object-level 3D generation. However, adapting 3D-GS from finite-scale 3D objects and humans to infinite-scale 3D cities is non-trivial. Unbounded 3D city generation entails significant storage overhead (out-of-memory issues), arising from the need to expand points to billions, often demanding hundreds of Gigabytes of VRAM for a city scene spanning 10km^2. In this paper, we propose **GaussianCity**, a generative Gaussian splatting framework dedicated to efficiently synthesizing unbounded 3D cities with a single feed-forward pass. Our key insights are two-fold: **1)** *Compact 3D Scene Representation*: We introduce BEV-Point as a highly compact intermediate representation, ensuring that the growth in VRAM usage for unbounded scenes remains constant, thus enabling unbounded city generation. **2)** *Spatial-aware Gaussian Attribute Decoder*: We present spatial-aware BEV-Point decoder to produce 3D Gaussian attributes, which leverages Point Serializer to integrate the structural and contextual characteristics of BEV points. Extensive experiments demonstrate that GaussianCity achieves state-of-the-art results in both drone-view and street-view 3D city generation. Notably, compared to CityDreamer, GaussianCity exhibits superior performance with a speedup of 60 times (10.72 FPS v.s. 0.18 FPS).
****

SVLTA: Benchmarking Vision-Language Temporal Alignment via Synthetic Video Situation
Hao Du, Bo Wu, Yan Lu, Zhendong Mao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13798-13809
Vision-language temporal alignment is a crucial capability for human dynamic recognition and cognition in real-world scenarios. While existing research focuses

on capturing vision-language relevance, it faces limitations due to biased tempo ral distributions, imprecise annotations, and insufficient compositionally. To a chieve fair evaluation and comprehensive exploration, our objective is to invest igate and evaluate the ability of models to achieve alignment from a temporal pe rspective, specifically focusing on their capacity to synchronize visual scenari os with linguistic context in a temporally coherent manner. As a preliminary ste p, we present the statistical analysis of existing benchmarks and reveal the exi sting challenges from a decomposed perspective. To this end, we introduce SVLTA, the synthetic vision-language temporal alignment derived via a well-designed an d feasible control generation method within a simulation environment. The approa ch considers commonsense knowledge, manipulable action, and constrained filterin g, which generates reasonable, diverse, and balanced data distributions for diag nostic evaluations. Our experiments reveal diagnostic insights through the evalu ations in temporal question answering, distributional shift sensitiveness, and t emporal alignment adaptation.

****

Mixture of Submodules for Domain Adaptive Person Search

Minsu Kim, Seungryong Kim, Kwanghoon Sohn; Proceedings of the Computer Vision an d Pattern Recognition Conference (CVPR), 2025, pp. 13990-14001

Existing technique on domain adaptive person search commonly utilizes the unifie d framework for jointly localizing and identifying the person across domains. Th is framework, however, inevitably results in the gradient conflict problem, part icularly in cross-domain scenarios with contradictory objectives, as the unified framework employs shared parameters to simultaneously address person detection and re-identification tasks across the domains. To overcome this, we present a n ovel mixture of submodules framework, dubbed MoS, that dynamically modulates the combination of submodules depending on the specific task to perform person dete ction and re-identification, separately. We further design the mixtures of submo dules that vary depending on the domain, enabling domain-specific knowledge tran sfer. Especially, we decompose the main model into several submodules and employ diverse mixtures of submodules that vary depending on the tasks and domains thr ough the conditional routing policy. In addition, we also present counterpart do main sample generation that synthesizes the augmented sample and uses them to le arn domain invariant representation for person re-identification through the con trastive domain alignment. We conduct experiments to demonstrate the effectivene ss of our MoS over the existing domain adaptive person search method and provide ablation studies.

****

Unsupervised Discovery of Facial Landmarks and Head Pose

Satyajit Tourani, Siddharth Tourani, Arif Mahmood, Muhammad Haris Khan; Proceedi ngs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21192-21202

Unsupervised landmark and head pose estimation is fundamental in fields like bio metrics, augmented reality, and emotion recognition, offering accurate spatial d ata without relying on labeled datasets. It enhances scalability, adaptability, and generalization across diverse settings, where manual labeling is costly. In this work we exploit Stable Diffusion to approach the challenging problem of uns upervised landmarks and head pose estimation and make following contributions. (a) We propose a semantic-aware landmark localization algorithm including a consi stent landmarks selection technique. (b) To encode landmarks and their holistic configuration, we propose learning image-aware textual embedding. (c) A novel al gorithm for landmarks-guided 3D head pose estimation is also proposed. (d) We re fine the landmarks using head pose by innovating a 3D rendering based augmentati on and pose-based batching technique while the refined landmarks, consequently i mproving the head pose. (e) We report a new state-of-the-art in unsupervised fa cial landmark estimation across five challenging datasets including AFLW2000, MA FL, Cat-Heads, LS3D and a facial landmark tracking benchmark 300VW. In unsupervi sed head pose estimation, we outperform existing methods on BIWI and AFLW2000 by visible margins. Moreover, our method provides a significant training speed-up over the existing best unsupervised landmark detection method.

****
Instruct-CLIP: Improving Instruction-Guided Image Editing with Automated Data Refinement Using Contrastive Learning
Sherry X. Chen, Misha Sra, Pradeep Sen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28513-28522
Although natural language instructions offer an intuitive way to guide automated image editing, deep-learning models often struggle to achieve high-quality results, largely due to the difficulty of creating large, high-quality training datasets. To do this, previous approaches have typically relied on text-to-image (T2I) generative models to produce pairs of original and edited images that simulate the input/output of an instruction-guided image- editing model. However, these image pairs often fail to align with the specified edit instructions due to the limitations of T2I models, which negatively impacts models trained on such datasets. To address this, we present Instruct-CLIP (I-CLIP), a selfsupervised method that learns the semantic changes between original and edited images to refine and better align the instructions in existing datasets. Furthermore, we adapt Instruct-CLIP to handle noisy latent images and diffusion timesteps so that it can be used to train latent diffusion models (LDMs) and efficiently enforce alignment between the edit instruction and the image changes in latent space at any step of the diffusion pipeline. We use Instruct-CLIP to correct the InstructPix2Pix dataset and get over 120K refined samples we then use to fine-tune their model, guided by our novel I- CLIP-based loss function. The resulting model can produce edits that are more aligned with the given instructions. Our code and dataset are available at https://github.com/SherryXTChen/Instruct-CLIP.git.
****
Stabilizing and Accelerating Autofocus with Expert Trajectory Regularized Deep Reinforcement Learning
Shouhang Zhu, Chenglin Li, Yuankun Jiang, Li Wei, Nuowen Kan, Ziyang Zheng, Wenrui Dai, Junni Zou, Hongkai Xiong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26440-26450
Autofocus is a crucial component of modern digital cameras. While recent learning-based methods achieve state-of-the-art in focus prediction accuracy, they unfortunately ignore the potential focus hunting phenomenon of back-and-forth lens movement in the multi-step focusing procedure. To address this, in this paper, we propose an expert regularized deep reinforcement learning (DRL)-based approach for autofocus, which can utilize the sequential information of lens movement trajectory to both enhance the multi-step in-focus prediction accuracy and reduce the chance of focus hunting. Our method generally follows an actor-critic framework. To accelerate the DRL's training with a higher sample efficiency, we initialize the policy with a pre-trained single-step prediction network, where the network is further improved by modifying the output of absolute in-focus position distribution to the relative lens movement distribution to establish a better mapping between input images and lens movement. To further stabilize DRL's training with a lower occurrence of focus hunting in the resulting lens movement trajectory, we generate some offline trajectories based on prior knowledge to avoid focus hunting, which are then leveraged as an offline dataset of expert trajectories to regularize the actor network's training. Empirical evaluations show that our method outperforms those learning-based methods on public benchmarks, with higher single- and multi-step prediction accuracy, and a significant reduction of focus hunting rate.
****
SharpDepth: Sharpening Metric Depth Predictions Using Diffusion Distillation
Duc-Hai Pham, Tung Do, Phong Nguyen, Binh-Son Hua, Khoi Nguyen, Rang Nguyen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17060-17069
We propose SharpDepth, a novel approach to monocular metric depth estimation that combines the metric accuracy of discriminative depth estimation methods (e.g., Metric3D, UniDepth) with the fine-grained boundary sharpness typically achieved by generative methods (e.g., Marigold, Lotus). Traditional discriminative models trained on real-world data with sparse ground-truth depth can accurately predi

ct metric depth but often produce over-smoothed or low-detail depth maps. Genera
tive models, in contrast, are trained on synthetic data with dense ground truth,
 generating depth maps with sharp boundaries yet only providing relative depth w
ith low accuracy. Our approach bridges these limitations by integrating metric a
ccuracy with detailed boundary preservation, resulting in depth predictions that
 are both metrically precise and visually sharp. Our extensive zero-shot evaluat
ions on standard depth estimation benchmarks confirm SharpDepth's effectiveness,
 showing its ability to achieve both high depth accuracy and detailed representa
tion, making it well-suited for applications requiring high-quality depth percep
tion across diverse, real-world environments.
****

Repurposing Stable Diffusion Attention for Training-Free Unsupervised Interactiv
e Segmentation

Markus Karmann, Onay Urfalioglu; Proceedings of the Computer Vision and Pattern
Recognition Conference (CVPR), 2025, pp. 24518-24528

Recent progress in interactive point prompt based Image Segmentation allows to s
ignificantly reduce the manual effort to obtain high quality semantic labels. St
ate-of-the-art unsupervised methods use self-supervised pre-trained models to ob
tain pseudo-labels which are used in training a prompt-based segmentation model.
 In this paper, we propose a novel unsupervised and training-free approach based
 solely on the self-attention of Stable Diffusion. We interpret the self-attenti
on tensor as a Markov transition operator, which enables us to iteratively const
ruct a Markov chain. Pixel-wise counting of the required number of iterations al
ong the Markov chain to reach a relative probability threshold yields a Markov-i
teration-map, which we simply call a Markov-map. Compared to the raw attention m
aps, we show that our proposed Markov-map has less noise, sharper semantic bound
aries and more uniform values within semantically similar regions. We integrate
the Markov-map in a simple yet effective truncated nearest neighbor framework to
 obtain interactive point prompt based segmentation. Despite being training-free
, we experimentally show that our approach yields excellent results in terms of
Number of Clicks (NoC), even outperforming state-of-the-art training based unsup
ervised methods in most of the datasets. Code is available at https://github.com
/mkarmann/m2n2.
****

GO-N3RDet: Geometry Optimized NeRF-enhanced 3D Object Detector

Zechuan Li, Hongshan Yu, Yihao Ding, Jinhao Qiao, Basim Azam, Naveed Akhtar; Pro
ceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025,
 pp. 27211-27221

We propose GO-N3RDet, a scene-geometry optimized multi-view 3D object detector e
nhanced by neural radiance fields. The key to accurate 3D object detection is in
 effective voxel representation. However, due to occlusion and lack of 3D inform
ation, constructing 3D features from multi-view 2D images is challenging. Addres
sing that, we introduce a unique 3D positional information embedded voxel optimi
zation mechanism to fuse multi-view features. To prioritize neural field reconst
ruction in object regions, we also devise a double importance sampling scheme fo
r the NeRF branch of our detector. We additionally propose an opacity optimizati
on module for precise voxel opacity prediction by enforcing multi-view consisten
cy constraints. Moreover, to further improve voxel density consistency across mu
ltiple perspectives, we incorporate ray distance as a weighting factor to minimi
ze cumulative ray errors. Our unique modules synergetically form an end-to-end n
eural model that establishes new state-of-the-art in NeRF-based multi-view 3D de
tection, verified with extensive experiments on ScanNet and ARKITScenes. Code wi
ll be available at https://github.com/ZechuanLi/GO-N3RDet.
****

DPSeg: Dual-Prompt Cost Volume Learning for Open-Vocabulary Semantic Segmentatio
n

Ziyu Zhao, Xiaoguang Li, Lingjia Shi, Nasrin Imanpour, Song Wang; Proceedings of
 the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25346-
25356

Open-vocabulary semantic segmentation aims to segment images into distinct seman

tic regions for both seen and unseen categories at the pixel level. Current methods utilize text embeddings from pre-trained vision-language models like CLIP but struggle with the inherent domain gap between image and text embeddings, even after extensive alignment during training. Additionally, relying solely on deep text-aligned features limits shallow-level feature guidance, which is crucial for detecting small objects and fine details, ultimately reducing segmentation accuracy. To address these limitations, we propose a dual prompting framework, DPSeg, for this task. Our approach combines dual-prompt cost volume generation, a cost volume-guided decoder, and a semantic-guided prompt refinement strategy that leverages our dual prompting scheme to mitigate alignment issues in visual prompt generation. By incorporating visual embeddings from a visual prompt encoder, our approach reduces the domain gap between text and image embeddings while providing multi-level guidance through shallow features. Extensive experiments demonstrate that our method significantly outperforms existing state-of-the-art approaches on multiple public datasets.

****

EvEnhancer: Empowering Effectiveness, Efficiency and Generalizability for Continuous Space-Time Video Super-Resolution with Events

Shuoyan Wei, Feng Li, Shengeng Tang, Yao Zhao, Huihui Bai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17755-17766

Continuous space-time video super-resolution (C-STVSR) endeavors to upscale videos simultaneously at arbitrary spatial and temporal scales, which has recently garnered increasing interest. However, prevailing methods struggle to yield satisfactory videos at out-of-distribution spatial and temporal scales. On the other hand, event streams characterized by high temporal resolution and high dynamic range, exhibit compelling promise in vision tasks. This paper presents EvEnhancer, an innovative approach that marries the unique advantages of event streams to elevate effectiveness, efficiency, and generalizability for C-STVSR. Our approach hinges on two pivotal components: 1) Event-adapted synthesis capitalizes on the spatiotemporal correlations between frames and events to discern and learn long-term motion trajectories, enabling the adaptive interpolation and fusion of informative spatiotemporal features; 2) Local implicit video transformer integrates local implicit video neural function with cross-scale spatiotemporal attention to learn continuous video representations utilized to generate plausible videos at arbitrary resolutions and frame rates. Experiments show that EvEnhancer achieves superiority on synthetic and real-world datasets and preferable generalizability on out-of-distribution scales against state-of-the-art methods. Code is available at https://github.com/W-Shuoyan/EvEnhancer.

****

Seeing A 3D World in A Grain of Sand

Yufan Zhang, Yu Ji, Yu Guo, Jinwei Ye; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11187-11196

We present a snapshot imaging technique for recovering 3D surrounding views of miniature scenes. Due to their intricacy, miniature scenes with objects sized in millimeters are difficult to reconstruct, yet miniatures are common in life and their 3D digitalization is desirable. We design a catadioptric imaging system with a single camera and eight pairs of planar mirrors for snapshot 3D reconstruction from a dollhouse perspective. We place paired mirrors on nested pyramid surfaces for capturing surrounding multi-view images in a single shot. Our mirror design is customizable based on the size of the scene for optimized view coverage. We use the 3D Gaussian Splatting (3DGS) representation for scene reconstruction and novel view synthesis. We overcome the challenge posed by our sparse view input by integrating visual hull-derived depth constraint. Our method demonstrates state-of-the-art performance on a variety of synthetic and real miniature scenes.

****

Simulator HC: Regression-based Online Simulation of Starting Problem-Solution Pairs for Homotopy Continuation in Geometric Vision

Xinyue Zhang, Zijia Dai, Wanting Xu, Laurent Kneip; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27103-27112

While automatically generated polynomial elimination templates have sparked great progress in the field of 3D computer vision, there remain many problems for which the degree of the constraints or the number of unknowns leads to intractability. In recent years, homotopy continuation has been introduced as a plausible alternative. However, the method currently depends on expensive parallel tracking of all possible solutions in the complex domain, or a classification network for starting problem-solution pairs trained over a limited set of real-world examples. Our innovation lies in a novel approach to finding solution-problem pairs, where we only need to predict a rough initial solution, with the corresponding problem generated by an online simulator. Subsequently, homotopy continuation is applied to track that single solution back to the original problem. We apply this elegant combination to generalized camera resectioning, and also introduce a new solution to the challenging generalized relative pose and scale problem. As demonstrated, the proposed method successfully compensates the raw error committed by the regressor alone, and leads to state-of-the-art efficiency and success rates.

****

Dynamic Integration of Task-Specific Adapters for Class Incremental Learning

Jiashuo Li, Shaokun Wang, Bo Qian, Yuhang He, Xing Wei, Qiang Wang, Yihong Gong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30545-30555

Non-exemplar Class Incremental Learning (NECIL) enables models to continuously acquire new classes without retraining from scratch and storing old task exemplars, addressing privacy and storage issues. However, the absence of data from earlier tasks exacerbates the challenge of catastrophic forgetting in NECIL. In this paper, we propose a novel framework called Dynamic Integration of task-specific Adapters (DIA), which comprises two key components: Task-Specific Adapter Integration (TSAI) and Patch-Level Model Alignment. TSAI boosts compositionality through a patch-level adapter integration strategy, aggregating richer task-specific information while maintaining low computation costs. Patch-Level Model Alignment maintains feature consistency and accurate decision boundaries via two specialized mechanisms: Patch-Level Distillation Loss (PDL) and Patch-Level Feature Reconstruction (PFR). Specifically, on the one hand, the PDL preserves feature-level consistency between successive models by implementing a distillation loss based on the contributions of patch tokens to new class learning. On the other hand, the PFR promotes classifier alignment by reconstructing old class features from previous tasks that adapt to new task knowledge, thereby preserving well-calibrated decision boundaries. Comprehensive experiments validate the effectiveness of our DIA, revealing significant improvements on NECIL benchmark datasets while maintaining an optimal balance between computational complexity and accuracy.

****

MoFlow: One-Step Flow Matching for Human Trajectory Forecasting via Implicit Maximum Likelihood Estimation based Distillation

Yuxiang Fu, Qi Yan, Lele Wang, Ke Li, Renjie Liao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17282-17293

In this paper, we address the problem of human trajectory forecasting, which aims to predict the inherently multi-modal future movements of humans based on their past trajectories and other contextual cues. We propose a novel motion prediction conditional flow matching model, termed MoFlow, to predict K-shot future trajectories for all agents in a given scene. We design a novel flow matching loss function that not only ensures at least one of the K sets of future trajectories is accurate but also encourages all K sets of future trajectories to be diverse and plausible. Furthermore, by leveraging the implicit maximum likelihood estimation (IMLE), we propose a novel distillation method for flow models that only requires samples from the teacher model. Extensive experiments on the real-world datasets, including SportVU NBA games, ETH-UCY, and SDD, demonstrate that both our teacher flow model and the IMLE-distilled student model achieve state-of-the-art performance. These models can generate diverse trajectories that are physically and socially plausible. Moreover, our one-step student model is 100 times faster than the teacher flow model during sampling. The code, model, and data are

available at our project page: https://moflow-imle.github.io/.
****
EgoPressure: A Dataset for Hand Pressure and Pose Estimation in Egocentric Visio
n
Yiming Zhao, Taein Kwon, Paul Streli, Marc Pollefeys, Christian Holz; Proceeding
s of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27
727-27738
Touch contact and pressure are essential for understanding how humans interact w
ith objects and offer insights that benefit applications in mixed reality and ro
botics. Estimating these interactions from an egocentric camera perspective is c
hallenging, largely due to the lack of comprehensive datasets that provide both
hand poses and pressure annotations. In this paper, we present EgoPressure, an e
gocentric dataset that is annotated with high-resolution pressure intensities at
contact points and precise hand pose meshes, obtained via our multi-view, seque
nce-based optimization method. We introduce baseline models for estimating appli
ed pressure on external surfaces from RGB images, both with and without hand pos
e information, as well as a joint model for predicting hand pose and the pressur
e distribution across the hand mesh.Our experiments show that pressure and hand
pose complement each other in understanding hand-object interactions.
****
DiverseFlow: Sample-Efficient Diverse Mode Coverage in Flows
Mashrur M. Morshed, Vishnu Boddeti; Proceedings of the Computer Vision and Patte
rn Recognition Conference (CVPR), 2025, pp. 23303-23312
Many real-world applications of flow-based generative models desire a diverse se
t of samples that cover multiple modes of the target distribution. However, the
predominant approach for obtaining diverse sets is not sample-efficient, as it i
nvolves independently obtaining many samples from the source distribution and ma
pping them through the flow until the desired mode coverage is achieved. As an a
lternative to repeated sampling, we introduce DiverseFlow: a training-free appro
ach to improve the diversity of flow models. Our key idea is to employ a determi
nantal point process to induce a coupling between the samples that drives divers
ity under a fixed sampling budget. In essence, DiverseFlow allows exploration of
more variations in a learned flow model with fewer samples. We demonstrate the
efficacy of our method for tasks where sample-efficient diversity is desirable,
such as text-guided image generation with polysemous words, inverse problems lik
e large-hole inpainting, and class-conditional image synthesis.
****
Reason-before-Retrieve: One-Stage Reflective Chain-of-Thoughts for Training-Free
Zero-Shot Composed Image Retrieval
Yuanmin Tang, Jue Zhang, Xiaoting Qin, Jing Yu, Gaopeng Gou, Gang Xiong, Qingwei
Lin, Saravan Rajmohan, Dongmei Zhang, Qi Wu; Proceedings of the Computer Vision
and Pattern Recognition Conference (CVPR), 2025, pp. 14400-14410
Composed Image Retrieval (CIR) aims to retrieve target images that closely resem
ble a reference image while integrating user-specified textual modifications, th
ereby capturing user intent more accurately. Existing training-free zero-shot CI
R (ZS-CIR) methods often employ a two-stage process: they first generate a capti
on for the reference image and then use Large Language Models for reasoning a ta
rget description. However, these methods suffer from missing critical visual det
ails and limited reasoning capabilities, leading to suboptimal retrieval perform
ance. To address these challenges, we propose a novel, training-free one-stage m
ethod, One-Stage Reflective Chain-of-Thought Reasoning (OSrCIR) for ZS-CIR, whic
h employs Multimodal Large Language Models to retain essential visual informatio
n in a single-stage reasoning process, eliminating the information loss in two-s
tage methods. Our Reflective Chain-of-Thought framework further improves interpr
etative accuracy by aligning manipulation intent with contextual cues from refer
ence images. OSrCIR achieves performance gains of 1.80% to 6.44% over existing t
raining-free methods across multiple tasks, setting new state-of-the-art results
in ZS-CIR and enhancing its utility in vision-language applications. Our code i
s available at https://github.com/microsoft/ACV/tree/main/OSrCIR.
****

UniGraspTransformer: Simplified Policy Distillation for Scalable Dexterous Robotic Grasping
Wenbo Wang, Fangyun Wei, Lei Zhou, Xi Chen, Lin Luo, Xiaohan Yi, Yizhong Zhang, Yaobo Liang, Chang Xu, Yan Lu, Jiaolong Yang, Baining Guo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12199-12208
We introduce UniGraspTransformer, a universal Transformer-based network for dexterous robotic grasping that simplifies training while enhancing scalability and performance. Unlike prior methods such as UniDexGrasp++, which require complex, multi-step training pipelines, UniGraspTransformer follows a streamlined process: first, dedicated policy networks are trained for individual objects using reinforcement learning to generate successful grasp trajectories; then, these trajectories are distilled into a single, universal network. Our approach enables UniGraspTransformer to scale effectively, incorporating up to 12 self-attention blocks for handling thousands of objects with diverse poses. Additionally, it generalizes well to both idealized and real-world inputs, evaluated in state-based and vision-based settings. Notably, UniGraspTransformer generates a broader range of grasping poses for objects in various shapes and orientations, resulting in more diverse grasp strategies. Experimental results demonstrate significant improvements over state-of-the-art, UniDexGrasp++, across various object categories, achieving success rate gains of 3.5%, 7.7%, and 10.1% on seen objects, unseen objects within seen categories, and completely unseen objects, respectively, in the vision-based setting. Project page: https://dexhand.github.io/UniGraspTransformer/
****
GeoMM: On Geodesic Perspective for Multi-modal Learning
Shibin Mei, Hang Wang, Bingbing Ni; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4776-4786
Geodesic distance serves as a reliable means of measuring distance in nonlinear spaces, and such nonlinear manifolds are prevalent in the current multimodal learning. In these scenarios, some samples may exhibit high similarity, yet they convey different semantics, making traditional distance metrics inadequate for distinguishing between positive and negative samples. This paper introduces geodesic distance as a novel distance metric in multi-modal learning for the first time, to mine correlations between samples, aiming to address the limitations of common distance metric. Our approach incorporates a comprehensive series of strategies to adapt geodesic distance for the current multimodal learning. Specifically, we construct a graph structure to represent the adjacency relationships among samples by thresholding distances between them and then apply the shortest-path algorithm to obtain geodesic distance within this graph. To facilitate efficient computation, we further propose a hierarchical graph structure through clustering and combined with incremental update strategies for dynamic status updates. Extensive experiments across various downstream tasks validate the effectiveness of our proposed method, demonstrating its capability to capture complex relationships between samples and improve the performance of multimodal learning models.
****
VISCO: Benchmarking Fine-Grained Critique and Correction Towards Self-Improvement in Visual Reasoning
Xueqing Wu, Yuheng Ding, Bingxuan Li, Pan Lu, Da Yin, Kai-Wei Chang, Nanyun Peng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9527-9537
The ability of large vision-language models (LVLMs) to critique and correct their reasoning is an essential building block towards their self-improvement. However, a systematic analysis of such capabilities in LVLMs is still lacking. We propose VISCO, the first benchmark to extensively analyze the fine-grained critique and correction capabilities of LVLMs. Compared to existing work that uses a single scalar value to critique the entire reasoning [4], VISCO features **dense** and **fine-grained** critique, requiring LVLMs to evaluate the correctness of each step in the chain-of-thought and provide natural language explanations to support their judgments. Extensive evaluation of 24 LVLMs demonstrates that human-written critiques significantly enhance the performance after correction, showcas

ing the potential of the self-improvement strategy. However, the model-generated critiques are less helpful and sometimes detrimental to the performance, suggesting that critique is the crucial bottleneck. We identified three common patterns in critique failures: failure to critique visual perception, reluctance to "say no", and exaggerated assumption of error propagation. To address these issues, we propose an effective LookBack strategy that revisits the image to verify each piece of information in the initial reasoning. \ourscritic significantly improves critique and correction performance by up to 13.5%.

****

MaskGWM: A Generalizable Driving World Model with Video Mask Reconstruction

Jingcheng Ni, Yuxin Guo, Yichen Liu, Rui Chen, Lewei Lu, Zehuan Wu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22381-22391

World models that forecast environmental changes from actions are vital for autonomous driving models with strong generalization. The prevailing driving world model mainly build on pixel-level video prediction model. Although these models can produce high-fidelity video sequences with advanced diffusion-based generator, they are constrained by their predictive duration and overall generalization capabilities. In this paper, we explore to solve this problem by combining pixel-level generation loss with MAE-style feature-level context learning. In particular, we instantiate this target with three key design: (1) A more scalable Diffusion Transformer (DiT) structure trained with extra mask construction task. (2) we devise diffusion-related mask tokens to deal with the fuzzy relations between mask reconstruction and generative diffusion process. (3) we extend mask construction task to spatial-temporal domain by utilizing row-wise mask for shifted self-attention rather than masked self-attention in MAE. Then, we adopt a row-wise cross-view module to align with this mask design. Based on above improvement, we propose MaskGWM: a Generalizable driving World Model embodied with Video Mask reconstruction. Our model contains two variants: MaskGWM-long, focusing on long-horizon prediction, and MaskGWM-mview, dedicated to multi-view generation.Comprehensive experiments on standard benchmarks validate the effectiveness of the proposed method, which contain normal validation of Nuscene dataset, long-horizon rollout of OpenDV-2K dataset and zero-shot validation of Waymo dataset. Quantitative metrics on these datasets show our method notably improving state-of-the-art driving world model.

****

3D-MVP: 3D Multiview Pretraining for Manipulation

Shengyi Qian, Kaichun Mo, Valts Blukis, David F. Fouhey, Dieter Fox, Ankit Goyal; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22530-22539

Recent works have shown that visual pretraining on egocentric datasets using masked autoencoders (MAE) can improve generalization for downstream robotics tasks. However, these approaches pretrain only on 2D images, while many robotics applications require 3D scene understanding. In this work, we propose 3D-MVP, a novel approach for 3D Multi-View Pretraining using masked autoencoders. We leverage Robotic View Transformer (RVT), which uses a multi-view transformer to understand the 3D scene and predict gripper pose actions. We split RVT's multi-view transformer into visual encoder and action decoder, and pretrain its visual encoder using masked autoencoding on large-scale 3D datasets such as Objaverse. We evaluate 3D-MVP on a suite of virtual robot manipulation tasks and demonstrate improved performance over baselines. Our results suggest that 3D-aware pretraining is a promising approach to improve generalization of vision-based robotic manipulation policies.

****

Enhanced OoD Detection through Cross-Modal Alignment of Multi-Modal Representations

Jeonghyeon Kim, Sangheum Hwang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29979-29988

Prior research on out-of-distribution detection (OoDD) has primarily focused on single-modality models. Recently, with the advent of large-scale pretrained visi

on-language models such as CLIP, OoDD methods utilizing such multi-modal representations through zero-shot and prompt learning strategies have emerged. However, these methods typically involve either freezing the pretrained weights or only partially tuning them, which can be suboptimal for downstream datasets. In this paper, we highlight that multi-modal fine-tuning (MMFT) can achieve notable OoDD performance. Despite some recent works demonstrating the impact of fine-tuning methods for OoDD, there remains significant potential for performance improvement. We investigate the limitation of naive fine-tuning methods, examining why they fail to fully leverage the pretrained knowledge. Our empirical analysis suggests that this issue could stem from the modality gap within in-distribution (ID) embeddings. To address this, we propose a training objective that enhances cross-modal alignment by regularizing the distances between image and text embeddings of ID data. This adjustment helps in better utilizing pretrained textual information by aligning similar semantics from different modalities (i.e., text and image) more closely in the hyperspherical representation space. We theoretically demonstrate that the proposed regularization corresponds to the maximum likelihood estimation of an energy-based model on a hypersphere. Utilizing ImageNet-1k OoD benchmark datasets, we show that our method, combined with post-hoc OoDD approaches leveraging pretrained knowledge (e.g., NegLabel), significantly outperforms existing methods, achieving state-of-the-art OoDD performance and leading ID accuracy.

****

Adaptive Dropout: Unleashing Dropout across Layers for Generalizable Image Super-Resolution

Hang Xu, Jie Huang, Wei Yu, Jiangtong Tan, Zhen Zou, Feng Zhao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7513-7523

Blind Super-Resolution(blind SR) aims to enhance the model's generalization ability with unknown degradation, yet it still encounters severe overfitting issues. Some previous methods inspired by dropout, which enhances generalization by regularizing features, have shown promising results in blind SR. Nevertheless, these methods focus solely on regularizing features before the final layer and overlook the need for generalization in features at intermediate layers. Without explicit regularization of features at intermediate layers, the blind SR network struggles to obtain well-generalized feature representations. However, the key challenge is that directly applying dropout to intermediate layers leads to a significant performance drop, which we attribute to the inconsistency in training-testing and across layers it introduced. Therefore, we propose Adaptive Dropout, a new regularization method for blind SR models, which mitigates the inconsistency and facilitates application across intermediate layers of networks. Specifically, for training-testing inconsistency, we re-design the form of dropout and integrate the features before and after dropout adaptively. For inconsistency in generalization requirements across different layers, we innovatively design an adaptive training strategy to strengthen feature propagation by layer-wise annealing. Experimental results show that our method outperforms all past regularization methods on both synthetic and real-world benchmark datasets, also highly effective in other image restoration tasks.

****

Breaking the Memory Barrier of Contrastive Loss via Tile-Based Strategy

Zesen Cheng, Hang Zhang, Kehan Li, Sicong Leng, Zhiqiang Hu, Fei Wu, Deli Zhao, Xin Li, Lidong Bing; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10036-10045

Contrastive loss is a powerful approach for representation learning, where larger batch sizes enhance performance by providing more negative samples to better distinguish between similar and dissimilar data. However, the full instantiation of the similarity matrix demands substantial GPU memory, making large batch training highly resource-intensive. To address this, we propose a tile-based computation strategy that partitions the contrastive loss calculation into small blocks, avoiding full materialization of the similarity matrix. Additionally, we introduce a multi-level tiling implementation to leverage the hierarchical structure

of distributed systems, using ring-based communication at the GPU level to optimize synchronization and fused kernels at the CUDA core level to reduce I/O overhead. Experimental results show that the proposed method significantly reduces GPU memory usage in contrastive loss. For instance, it enables contrastive training of a CLIP-ViT-L/14 model with a batch size of 4M using only 8 A800 80GB GPUs, without sacrificing accuracy. Compared to state-of-the-art memory-efficient solutions, it achieves a two-order-of-magnitude reduction in memory while maintaining comparable speed. The code will be made publicly available.
****

Mimir: Improving Video Diffusion Models for Precise Text Understanding
Shuai Tan, Biao Gong, Yutong Feng, Kecheng Zheng, Dandan Zheng, Shuwei Shi, Yujun Shen, Jingdong Chen, Ming Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23978-23988
Text serves as the key control signal in video generation due to its narrative nature. To render text descriptions into video clips, current video diffusion models borrow features from text encoders yet struggle with limited text comprehension. The recent success of large language models (LLMs) showcases the power of decoder-only transformers, which offers three clear benefits for text-to-video (T2V) generation, namely, precise text understanding resulting from the superior scalability, imagination beyond the input text enabled by next token prediction, and flexibility to prioritize user interests through instruction tuning. Nevertheless, the feature distribution gap emerging from the two different text modeling paradigms hinders the direct use of LLMs in established T2V models. This work addresses this challenge with Mimir, an end-to-end training framework featuring a carefully tailored token fuser to harmonize the outputs from text encoders and LLMs. Such a design allows the T2V model to fully leverage learned video priors while capitalizing on the text-related capability of LLMs. Extensive quantitative and qualitative results demonstrate the effectiveness of our approach in generating high-quality videos with excellent text comprehension, especially when processing short captions and managing shifting motions. The code and models will be made publicly available.
****

UCM-VeID V2: A Richer Dataset and A Pre-training Method for UAV Cross-Modality Vehicle Re-Identification
Xingyue Liu, Jiahao Qi, Chen Chen, KangCheng Bin, Ping Zhong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22286-22295
Cross-Modality Re-Identification (VI-ReID) aims to achieve around-the-clock target matching, benefiting from the strengths of both RGB and infrared (IR) modalities. However, the field is hindered by limited datasets, particularly for vehicle VI-ReID, and by challenges such as modality bias training (MBT), stemming from biased pre-training on ImageNet. To tackle the above issues, this paper introduces an UCM-VeID V2 dataset benchmark for vehicle VI-ReID, and proposes a new self-supervised pre-training method, Cross-Modality Patch-Mixed Self-supervised Learning (PMSL). UCM-VeID V2 dataset features a significant increase in data volume, along with enhancements in multiple aspects. PMSL addresses MBT by learning modality-invariant features through Patch-Mixed Image Reconstruction (PMIR) and Modality Discrimination Adversarial Learning (MDAL), and enhances discriminability with Modality-Augmented Contrasting Cluster (MACC). Comprehensive experiments are carried out to validate the effectiveness of the proposed method.
****

Learning Phase Distortion with Selective State Space Models for Video Turbulence Mitigation
Xingguang Zhang, Nicholas Chimitt, Xijun Wang, Yu Yuan, Stanley H. Chan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2127-2138
Atmospheric turbulence is a major source of image degradation in long-range imaging systems. Although numerous deep learning-based turbulence mitigation (TM) methods have been proposed, many are slow, memory-hungry, and do not generalize well. In the spatial domain, methods based on convolutional operators have a limit

ed receptive field, so they cannot handle a large spatial dependency required by turbulence. In the temporal domain, methods relying on self-attention can, in theory, leverage the lucky effects of turbulence, but their quadratic complexity makes it difficult to scale to many frames. Traditional recurrent aggregation methods face parallelization challenges. In this paper, we present a new TM method based on two concepts: (1) A turbulence mitigation network based on the Selective State Space Model (MambaTM). MambaTM provides a global receptive field in each layer across spatial and temporal dimensions while maintaining linear computational complexity. (2) Learned Latent Phase Distortion (LPD). LPD guides the state space model. Unlike classical Zernike-based representations of phase distortion, the new LPD map uniquely captures the actual effects of turbulence, significantly improving the model's capability to estimate degradation by reducing the ill-posedness. Our proposed method exceeds current state-of-the-art networks on various synthetic and real-world TM benchmarks with significantly faster inference speed. The code will be publicly available.

****

RoboPEPP: Vision-Based Robot Pose and Joint Angle Estimation through Embedding Predictive Pre-Training

Raktim Gautam Goswami, Prashanth Krishnamurthy, Yann LeCun, Farshad Khorrami; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6930-6939

Vision-based pose estimation of articulated robots with unknown joint angles has applications in collaborative robotics and human-robot interaction tasks. Current frameworks use neural network encoders to extract image features and downstream layers to predict joint angles and robot pose. While images of robots inherently contain rich information about the robot's physical structures, existing methods often fail to leverage it fully; therefore, limiting performance under occlusions and truncations. To address this, we introduce RoboPEPP, a method that fuses information about the robot's physical model into the encoder using a masking-based self-supervised embedding-predictive architecture. Specifically, we mask the robot's joints and pre-train an encoder-predictor model to infer the joints' embeddings from surrounding unmasked regions, enhancing the encoder's understanding of the robot's physical model. The pre-trained encoder-predictor pair, along with joint angle and keypoint prediction networks, is then fine-tuned for pose and joint angle estimation. Random masking of input during fine-tuning and keypoint filtering during evaluation further improves robustness. Our method, evaluated on several datasets, achieves the best results in robot pose and joint angle estimation while being the least sensitive to occlusions and requiring the lowest execution time.

****

Distraction is All You Need for Multimodal Large Language Model Jailbreaking

Zuopeng Yang, Jiluan Fan, Anli Yan, Erdun Gao, Xin Lin, Tao Li, Kanghua Mo, Changyu Dong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9467-9476

Multimodal Large Language Models (MLLMs) bridge the gap between visual and textual data, enabling a range of advanced applications. However, complex internal interactions among visual elements and their alignment with text can introduce vulnerabilities, which may be exploited to bypass safety mechanisms. To address this, we analyze the relationship between image content and task and find that the complexity of subimages, rather than their content, is key. Building on this insight, we propose the Distraction Hypothesis, followed by a novel framework called Contrasting Subimage Distraction Jailbreaking (CS-DJ), to achieve jailbreaking by disrupting MLLMs alignment through multi-level distraction strategies. CS-DJ consists of two components: structured distraction, achieved through query decomposition that induces a distributional shift by fragmenting harmful prompts into sub-queries, and visual-enhanced distraction, realized by constructing contrasting subimages to disrupt the interactions among visual elements within the model. This dual strategy disperses the model's attention, reducing its ability to detect and mitigate harmful content. Extensive experiments across five representative scenarios and four popular closed-source MLLMs, including \texttt GPT-4o-mi

ni , \texttt GPT-4o , \texttt GPT-4V , and \texttt Gemini-1.5-Flash , demonstrat e that CS-DJ achieves average success rates of 52.40% for the attack success rat e and 74.10% for the ensemble attack success rate. These results reveal the pote ntial of distraction-based approaches to exploit and bypass MLLMs' defenses, off ering new insights for attack strategies.
****

Apollo: An Exploration of Video Understanding in Large Multimodal Models
Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-E struch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, Serena Yeung-Lev y, Xide Xia; Proceedings of the Computer Vision and Pattern Recognition Conferen ce (CVPR), 2025, pp. 18891-18901
Despite the rapid integration of video perception capabilities into Large Multim odal Models (LMMs), what drives their video perception remains poorly understood . Consequently, many design decisions in this domain are made without proper jus tification or analysis. The high computational cost of training and evaluating s uch models and limited open research hinder the development of video-LMMs. To ad dress this, we present a comprehensive study that helps uncover what effectively drives video understanding in LMMs. We begin by critically examining the primar y contributors to the high computational requirements associated with video-LMM research and discover Scaling Consistency, wherein design and training decisions made on smaller models and datasets (up to a critical size) effectively transfe r to larger models. Leveraging these insights, we explored many video-specific a spects of video-LMMs, including video sampling, architectures, data composition, training schedules, and more. Guided by these findings, we introduce Apollo, a state-of-the-art family of LMMs that achieve superior performance across differe nt model sizes. Our models process over 1-hour videos efficiently, with the 3B p arameter variant outperforming most existing 7B models. Apollo-7B is state-of-th e-art compared to 7B LMMs with a 70.9 on MLVU, and 63.3 on Video-MME.
****

Skip Tuning: Pre-trained Vision-Language Models are Effective and Efficient Adap ters Themselves
Shihan Wu, Ji Zhang, Pengpeng Zeng, Lianli Gao, Jingkuan Song, Heng Tao Shen; Pr oceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025 , pp. 14723-14732
Prompt tuning (PT) has long been recognized as an effective and efficient paradi gm for transferring large pre-trained vision-language models (VLMs) to downstrea m tasks by learning a tiny set of context vectors. Nevertheless, in this work, w e reveal that freezing the parameters of VLMs during learning the context vector s neither facilitates the transferability of pre-trained knowledge nor improves the memory and time efficiency significantly. Upon further investigation, we fin d that reducing both the length and width of the feature-gradient propagation fl ows of the full fine-tuning (FT) baseline is key to achieving effective and effi cient knowledge transfer. Motivated by this, we propose Skip Tuning, a novel par adigm for adapting VLMs to downstream tasks. Unlike existing PT or adapter-based methods, Skip Tuning applies Layer-wise Skipping (LSkip) and Class-wise Skippin g (CSkip) upon the FT baseline without introducing extra context vectors or adap ter modules. Extensive experiments across a wide spectrum of benchmarks demonstr ate the superior effectiveness and efficiency of our Skip Tuning over both PT an d adapter-based methods. Code: https://github.com/anonymity-007/SkipT.
****

PatchDPO: Patch-level DPO for Finetuning-free Personalized Image Generation
Qihan Huang, Long Chan, Jinlong Liu, Wanggui He, Hao Jiang, Mingli Song, Jie Son g; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18369-18378
Finetuning-free personalized image generation can synthesize customized images w ithout test-time finetuning, attracting wide research interest owing to its high efficiency. Current finetuning-free methods simply adopt a single training stag e with a simple image reconstruction task, and they typically generate low-quali ty images inconsistent with the reference images during test-time. To mitigate t his problem, inspired by the recent DPO (i.e., direct preference optimization) t

echnique, this work proposes an additional training stage to improve the pre-tra
ined personalized generation models. However, traditional DPO only determines th
e overall superiority or inferiority of two samples, which is not suitable for p
ersonalized image generation because the generated images are commonly inconsist
ent with the reference images only in some local image patches. To tackle this p
roblem, this work proposes PatchDPO that estimates the quality of image patches
within each generated image and accordingly trains the model. To this end, Patch
DPO first leverages the pre-trained vision models with a proposed self-supervise
d training method to estimate the patch quality. Next, PatchDPO adopts a weighte
d training approach to train the model with the estimated patch quality, which r
ewards the image patches with high quality while penalizing the image patches wi
th low quality. Experiment results demonstrate that PatchDPO significantly impro
ves the performance of multiple pre-trained personalized generation models, and
achieves state-of-the-art performance on both single-object and multi-object per
sonalized image generation. Our code is available at https://github.com/hqhQAQ/P
atchDPO.
****

Learning to Normalize on the SPD Manifold under Bures-Wasserstein Geometry
Rui Wang, Shaocheng Jin, Ziheng Chen, Xiaoqing Luo, Xiao-Jun Wu; Proceedings of
the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8289-82
98
Covariance matrices have proven highly effective across many scientific fields.
Since these matrices lie within the Symmetric Positive Definite (SPD) manifold--
a Riemannian space with intrinsic non-Euclidean geometry, the primary challenge
in representation learning is to respect this underlying geometric structure. Dr
awing inspiration from the success of Euclidean deep learning, researchers have
developed neural networks on the SPD manifolds for more faithful covariance embe
dding learning. A notable advancement in this area is the implementation of Riem
annian batch normalization (RBN), which has been shown to improve the performanc
e of SPD network models. Nonetheless, the Riemannian metric beneath the existing
 RBN might fail to effectively deal with the ill-conditioned SPD matrices (ICSM)
, undermining the effectiveness of RBN. In contrast, the Bures-Wasserstein metri
c (BWM) demonstrates superior performance for ill-conditioning. In addition, the
 recently introduced Generalized BWM (GBWM) parameterizes the vanilla BWM via an
 SPD matrix, allowing for a more nuanced representation of vibrant geometries of
 the SPD manifold. Therefore, we propose a novel RBN algorithm based on the GBW
geometry, incorporating a learnable metric parameter. Moreover, the deformation
of GBWM by matrix power is also introduced to further enhance the representation
al capacity of GBWM-based RBN. Experimental results on different datasets valida
te the effectiveness of our proposed method. The code is available at https://gi
thub.com/jjscc/GBWBN.
****

SAMWISE: Infusing Wisdom in SAM2 for Text-Driven Video Segmentation
Claudia Cuttano, Gabriele Trivigno, Gabriele Rosi, Carlo Masone, Giuseppe Averta
; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR),
2025, pp. 3395-3405
Referring Video Object Segmentation (RVOS) relies on natural language expression
s to segment an object in a video clip. Existing methods restrict reasoning eith
er to independent short clips, losing global context, or process the entire vide
o offline, impairing their application in a streaming fashion. In this work, we
aim to surpass these limitations and design an RVOS method capable of effectivel
y operating in streaming-like scenarios while retaining contextual information f
rom past frames. We build upon the Segment-Anything 2 (SAM2) model, that provide
s robust segmentation and tracking capabilities and is naturally suited for stre
aming processing. We make SAM2 wiser, by empowering it with natural language und
erstanding and explicit temporal modeling at the feature extraction stage, with
out fine-tuning its weights, and without outsourcing modality interaction to ext
ernal models. To this end, we introduce a novel adapter module that injects temp
oral information and multi-modal cues in the feature extraction process. We furt
her reveal the phenomenon of tracking bias in SAM2 and propose a learnable modul

e to adjust its tracking focus when the current frame features suggest a new obj ect more aligned with the caption. Our proposed method, SAMWISE, achieves state- of-the-art across various benchmarks, by adding a negligible overhead of less th an 5 M parameters. Code is available at https://github.com/ClaudiaCuttano/SAMWIS E.
****

MegaSaM: Accurate, Fast and Robust Structure and Motion from Casual Dynamic Vide os

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, Noah Snavely; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10486-10496

We present a system that allows for accurate, fast, and robust estimation of cam era parameters and depth maps from casual monocular videos of dynamic scenes. Mo st conventional structure from motion and monocular SLAM techniques assume input videos that feature predominantly static scenes with large amounts of parallax. Such methods tend to produce erroneous estimates in the absence of these condit ions. Recent neural network based approaches attempt to overcome these challenge s; however, such methods are either computationally expensive or brittle when ru n on dynamic videos with uncontrolled camera motion or unknown field of view. We demonstrate the surprising effectiveness of the deep visual SLAM framework, and with careful modifications to its training and inference schemes, this system c an scale to real-world videos of complex dynamic scenes with unconstrained camer a paths, including videos with little camera parallax. Extensive experiments on both synthetic and real videos demonstrate that our system is significantly more accurate and robust at camera pose and depth estimation when compared with prio r and concurrent work, with faster or comparable running times.
****

BEVDiffuser: Plug-and-Play Diffusion Model for BEV Denoising with Ground-Truth G uidance

Xin Ye, Burhaneddin Yaman, Sheng Cheng, Feng Tao, Abhirup Mallik, Liu Ren; Proce edings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, p p. 1495-1504

Bird's-eye-view (BEV) representations play a crucial role in autonomous driving tasks. Despite recent advancements in BEV generation, inherent noise, stemming f rom sensor limitations and the learning process, remains largely unaddressed, re sulting in suboptimal BEV representations that adversely impact the performance of downstream tasks. To address this, we propose BEVDiffuser, a novel diffusion model that effectively denoises BEV feature maps using the ground-truth object l ayout as guidance. BEVDiffuser can be operated in a plug-and-play manner during training time to enhance existing BEV models without requiring any architectura l modifications. Extensive experiments on the challenging nuScenes dataset demon strate BEVDiffuser's exceptional denoising and generation capabilities, which en able significant enhancement to existing BEV models, as evidenced by notable imp rovements of 12.3% in mAP and 10.1% in NDS achieved for 3D object detection with out introducing additional computational complexity. Moreover, substantial impro vements in long-tail object detection and under challenging weather and lighting conditions further validate BEVDiffuser's effectiveness in denoising and enhanc ing BEV representations.
****

GeoAvatar: Geometrically-Consistent Multi-Person Avatar Reconstruction from Spar se Multi-View Videos

Soohyun Lee, Seoyeon Kim, HeeKyung Lee, Won-Sik Jeong, Joo Ho Lee; Proceedings o f the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21138 -21147

Multi-person avatar reconstruction from sparse multi-view videos is challenging. The independent avatar reconstruction of each person often fails to reconstruct the geometric relationship among multiple instances, resulting in inter-penetra tions among avatars. Some researchers resolve this issue via neural volumetric r endering techniques but they suffer from huge computational costs for rendering and training. In this paper, we propose a multi-person avatar reconstruction met

hod that reconstructs a 3D avatar of each person while keeping the geometric rel ations among people. Our 2D Gaussian Splatting (2DGS)-based avatar representatio n allows us to represent geometrically-accurate surfaces of multiple instances t hat support sharp inside-outside tests. We utilize the monocular prior to allevi ate the inter-penetration via surface ordering and to enhance the geometry in le ss-observed and textureless surfaces. We demonstrate the efficiency and performa nce of our method quantitatively and qualitatively on a multi-person dataset con taining close interactions.
****

FinePhys: Fine-grained Human Action Generation by Explicitly Incorporating Physi cal Laws for Effective Skeletal Guidance
Dian Shao, Mingfei Shi, Shengda Xu, Haodong Chen, Yongle Huang, Binglu Wang; Pro ceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1905-1916
Although remarkable progress has been achieved in video generation, synthesizing physically plausible human actions remains an unresolved challenge, especially when addressing fine-grained semantics and complex temporal dynamics. For instan ce, generating gymnastics routines such as "two turns on one leg with the free l eg optionally below horizontal" poses substantial difficulties for current video generation methods, which often fail to produce satisfactory results. To addres s this, we propose FinePhys, a Fine-grained human action generation framework in corporating Physics for effective skeletal guidance. Specifically, FinePhys firs t performs online 2D pose estimation and then accomplishes dimension lifting thr ough in-context learning. Recognizing that such data-driven 3D pose estimations may lack stability and interpretability, we incorporate a physics-based module t hat re-estimates motion dynamics using Euler-Lagrange equations, calculating joi nt accelerations bidirectionally across the temporal dimension. The physically p redicted 3D poses are then fused with data-driven poses to provide multi-scale 2 D heatmap-based guidance for the video generation process. Evaluated on three fi ne-grained action subsets from FineGym (FX-JUMP, FX-TURN, and FX-SALTO), FinePhy s significantly outperforms competitive baselines. Comprehensive qualitative res ults further demonstrate FinePhys's ability to generate more natural and plausib le fine-grained human actions.
****

DiET-GS: Diffusion Prior and Event Stream-Assisted Motion Deblurring 3D Gaussian Splatting
Seungjun Lee, Gim Hee Lee; Proceedings of the Computer Vision and Pattern Recogn ition Conference (CVPR), 2025, pp. 21739-21749
Reconstructing sharp 3D representations from blurry multi-view images are long-s tanding problem in computer vision. Recent works attempt to enhance high-quality novel view synthesis from the motion blur by leveraging event-based cameras, be nefiting from high dynamic range and microsecond temporal resolution. However, t hey often reach sub-optimal visual quality in either restoring inaccurate color or losing fine-grained details. In this paper, we present DiET-GS, a diffusion p rior and event stream-assisted motion deblurring 3DGS. Our framework effectively leverages blur-free event streams and diffusion prior in a two-stage training s trategy. Specifically, we introduce the novel framework to constraint 3DGS with event double integral, achieving both accurate color and well-defined details. A dditionally, we propose a simple technique to leverage diffusion prior to furthe r enhance the edge details. Qualitative and quantitative results on both synthet ic and real-world data demonstrate that our DiET-GS is capable of producing bett er quality of novel views compared to the existing baselines. The project page l ink is attached in main paper.
****

Speedy-Splat: Fast 3D Gaussian Splatting with Sparse Pixels and Sparse Primitive s
Alex Hanson, Allen Tu, Geng Lin, Vasu Singla, Matthias Zwicker, Tom Goldstein; P roceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 202 5, pp. 21537-21546
3D Gaussian Splatting (3D-GS) is a recent 3D scene reconstruction technique that

enables real-time rendering of novel views by modeling scenes as parametric point clouds of differentiable 3D Gaussians. However, its rendering speed and model size still present bottlenecks, especially in resource-constrained settings. In this paper, we identify and address two key inefficiencies in 3D-GS to substantially improve rendering speed.These improvements also yield the ancillary benefits of reduced model size and training time. First, we optimize the rendering pipeline to precisely localize Gaussians in the scene, boosting rendering speed without altering visual fidelity. Second, we introduce a novel pruning technique and integrate it into the training pipeline, significantly reducing model size and training time while further raising rendering speed. Our Speedy-Splat approach combines these techniques to accelerate average rendering speed by a drastic 6.71x across scenes from the Mip-NeRF 360, Tanks&Temples, and Deep Blending datasets.

****

SeedVR: Seeding Infinity in Diffusion Transformer Towards Generic Video Restoration

Jianyi Wang, Zhijie Lin, Meng Wei, Yang Zhao, Ceyuan Yang, Chen Change Loy, Lu Jiang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2161-2172

Video restoration poses non-trivial challenges in maintaining fidelity while recovering temporally consistent details from unknown degradations in the wild. Despite recent advances in diffusion-based restoration, these methods often face limitations in generation capability and sampling efficiency. In this work, we present \text SeedVR , a diffusion transformer designed to handle real-world video restoration with arbitrary length and resolution. The core design of SeedVR lies in the shifted window attention that facilitates effective restoration on long video sequences. SeedVR further supports variable-sized windows near the boundary of both spatial and temporal dimensions, overcoming the resolution constraints of traditional window attention. Equipped with contemporary practices, including causal video autoencoder, mixed image and video training, and progressive training, SeedVR achieves highly-competitive performance on both synthetic and real-world benchmarks, as well as AI-generated videos. Extensive experiments demonstrate SeedVR's superiority over existing methods for generic video restoration.

****

Beyond Generation: A Diffusion-based Low-level Feature Extractor for Detecting AI-generated Images

Nan Zhong, Haoyu Chen, Yiran Xu, Zhenxing Qian, Xinpeng Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8258-8268

The prevalence of AI-generated images has evoked concerns regarding the potential misuse of image generation technologies. In response, numerous detection methods aim to identify AI-generated images by analyzing generative artifacts. Unfortunately, most detectors quickly become obsolete with the development of generative models. In this paper, we first design a low-level feature extractor that transforms spatial images into feature space, where different source images exhibit distinct distributions. The pretext task for the feature extractor is to distinguish between images that differ only at the pixel level. This image set comprises the original image as well as versions that have been subjected to varying levels of noise and subsequently denoised using a pre-trained diffusion model. We employ the diffusion model as a denoising tool rather than an image generation tool. Then, we frame the AI-generated image detection task as a one-class classification. We estimate the low-level intrinsic feature distribution of real photographic images and identify features that deviate from this distribution as indicators of AI-generated images. We evaluate our method against over 20 different generative models, including those in GenImage and DRCT-2M datasets. Extensive experiments demonstrate its effectiveness on AI-generated images produced not only by diffusion models but also by GANs, flow-based models, and their variants.

****

Robust-MVTON: Learning Cross-Pose Feature Alignment and Fusion for Robust Multi-View Virtual Try-On

Nannan Zhang, Yijiang Li, Dong Du, Zheng Chong, Zhengwentai Sun, Jianhao Zeng, Y

usheng Dai, Zhengyu Xie, Hairui Zhu, Xiaoguang Han; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16029-16039
This paper tackles the emerging challenge of multi-view virtual try-on, utilizing both front- and back-view clothing images as inputs. Extending frontal try-on methods to a multi-view context is not straightforward. Simply concatenating the two input views or encoding their features for a generative model, such as a diffusion model, often fails to produce satisfactory results. The main challenge lies in effectively extracting and fusing meaningful clothing features from these input views. Existing explicit warping based methods, which establish direct correspondence between input and target views, tend to introduce artifacts, particularly when there is a significant disparity between the input and target views. Conversely, implicit encoding based methods often lose spatial information about clothing, resulting in outputs that lack detail. To overcome these challenges, we propose Robust-MVTON, an end-to end method for robust and high-quality multi-view try-ons. Our approach introduces a novel cross-pose feature alignment technique to guide the fusion of clothing features and incorporates a newly designed loss function for training. With the fused multi-scale clothing features, we employ a coarse-to-fine diffusion model to generate realistic and detailed results. Extensive experiments conducted on the Deepfashion and MPV datasets affirm the superiority of our method, achieving state-of-the-art performance.
****

Omnia de EgoTempo: Benchmarking Temporal Understanding of Multi-Modal LLMs in Egocentric Videos
Chiara Plizzari, Alessio Tonioni, Yongqin Xian, Achin Kulshrestha, Federico Tombari; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24129-24138
Understanding fine-grained temporal dynamics is crucial in egocentric videos, where continuous streams capture frequent, close-up interactions with objects. In this work, we bring to light that current egocentric video question-answering datasets often include questions that can be answered using only few frames or commonsense reasoning, without being necessarily grounded in the actual video. Our analysis shows that state-of-the-art Multi-Modal Large Language Models (MLLMs) on these benchmarks achieve remarkably high performance using just text or a single frame as input. To address these limitations, we introduce EgoTempo, a dataset specifically designed to evaluate temporal understanding in the egocentric domain. EgoTempo emphasizes tasks that require integrating information across the entire video, ensuring that models would need to rely on temporal patterns rather than static cues or pre-existing knowledge. Extensive experiments on EgoTempo show that current MLLMs still fall short in temporal reasoning on egocentric videos, and thus we hope EgoTempo will catalyze new research in the field and inspire models that better capture the complexity of temporal dynamics. Dataset and code are available at https://github.com/google-research-datasets/egotempo.git.
****

ODHSR: Online Dense 3D Reconstruction of Humans and Scenes from Monocular Videos
Zetong Zhang, Manuel Kaufmann, Lixin Xue, Jie Song, Martin R. Oswald; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21824-21835
Creating a photorealistic scene and human reconstruction from a single monocular in-the-wild video figures prominently in the perception of a human-centric 3D world. Recent neural rendering advances have enabled holistic human-scene reconstruction but require pre-calibrated camera and human poses, and days of training time. In this work, we introduce a novel unified framework that simultaneously performs camera tracking, human pose estimation and human-scene reconstruction in an online fashion. 3D Gaussian Splatting is utilized to learn Gaussian primitives for humans and scenes efficiently, and reconstruction-based camera tracking and human pose estimation modules are designed to enable holistic understanding and effective disentanglement of pose and appearance. Specifically, we design a human deformation module to reconstruct the details and enhance generalizability to out-of-distribution poses faithfully. Aiming to learn the spatial correlation between human and scene accurately, we introduce occlusion-aware human silhouet

te rendering and monocular geometric priors, which further improve reconstructio n quality. Experiments on the EMDB and NeuMan datasets demonstrate superior or o n-par performance with existing methods in camera tracking, human pose estimatio n, novel view synthesis and runtime. Our project page is at https://eth-ait.gith ub.io/ODHSR.
****

Identity-Preserving Text-to-Video Generation by Frequency Decomposition
Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, Li Yuan; Proceedings of the Computer Vision and Pattern Recognition Confer ence (CVPR), 2025, pp. 12978-12988
Identity-preserving text-to-video (IPT2V) generation aims to create high-fidelit y videos with consistent human identity. It is an important task in video gener ation but remains an open problem for generative models. This paper pushes the t echnical frontier of IPT2V in two directions that have not been resolved in the literature: (1) A tuning-free pipeline without tedious case-by-case finetuning, and (2) A frequency-aware heuristic identity-preserving Diffusion Transformer (D iT)-based control scheme. To achieve these goals, we propose ConsisID, a tuning-free DiT-based controllable IPT2V model to keep human-identity consistent in the generated video. Inspired by prior findings in frequency analysis of vision/dif fusion transformers, it employs identity-control signals base on frequency domai n, since facial features can be decomposed into low-frequency global features (e .g., profile, proportions) and high-frequency intrinsic features (e.g., identity markers that remain unaffected by pose changes). Extensive experiments demonstr ate that our frequency-aware heuristic scheme provides an optimal control soluti on for DiT-based models, making strides towards more effective IPT2V.
****

FreeGave: 3D Physics Learning from Dynamic Videos by Gaussian Velocity
Jinxi Li, Ziyang Song, Siyuan Zhou, Bo Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12433-12443
In this paper, we aim to model 3D scene geometry, appearance, and the underlying physics purely from multi-view videos. By applying various governing PDEs as PI NN losses or incorporating physics simulation into neural networks, existing wor ks often fail to learn complex physical motions at boundaries or require object priors such as masks or types. In this paper, we propose FreeGave to learn physi cs of complex dynamic 3D scenes without needing any object priors. The key to ou r approach is to introduce a physics code followed by a carefully designed diver gence-free module for estimating a per-Gaussian velocity field, without relying on the inefficient PINN losses. Extensive experiments on three public datasets a nd a newly collected challenging real-world dataset demonstrate the superior per formance of our method for future frame extrapolation and motion segmentation. M ost notably, our investigation into the learned physics codes reveals that they truly learn meaningful 3D physical motion patterns in the absence of any human l abels in training. Our code and data are available at https://github.com/vLAR-gr oup/FreeGave.
****

SpiritSight Agent: Advanced GUI Agent with One Look
Zhiyuan Huang, Ziming Cheng, Junting Pan, Zhaohui Hou, Mingjie Zhan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 294 90-29500
Graphical User Interface (GUI) agents demonstrate promising potential in assisti ng human-computer interaction, automating human user's navigation on digital dev ices. An ideal GUI agent is expected to achieve high accuracy, low latency, and compatibility for different GUI platforms. Recent vision-based approaches have s hown promise by leveraging advanced Vision Language Models (VLMs). While they ge nerally meet the requirements of compatibility and low latency, these vision-bas ed GUI agents tend to have low accuracy due to their limitations in element grou nding. To address this issue, we propose SpiritSight, a vision-based, end-to-end GUI agent that excels in GUI navigation tasks across various GUI platforms. Fir st, we create a multi-level, large-scale, high-quality GUI dataset called GUI-La sagne using scalable methods, empowering SpiritSight with robust GUI understandi

ng and grounding capabilities. Second, we introduce the Universal Block Parsing (UBP) method to resolve the ambiguity problem inherited from the dynamic resolution strategy, further enhancing SpiritSight's ability to ground GUI objects. Through these efforts, SpiritSight agent outperforms other advanced methods on diverse GUI benchmarks, demonstrating its superior capability and compatibility in GUI navigation tasks. The models and code will be made available upon publication.

****

Zero-Shot Monocular Scene Flow Estimation in the Wild

Yiqing Liang, Abhishek Badki, Hang Su, James Tompkin, Orazio Gallo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21031-21044

Large models have shown generalization across datasets for many low-level vision tasks, like depth estimation, but no such general models exist for scene flow. Even though scene flow prediction has wide potential, its practical use is limited because of the lack of generalization of current predictive models. We identify three key challenges and propose solutions for each. First, we create a method that jointly estimates geometry and motion for accurate prediction. Second, we alleviate scene flow data scarcity with a data recipe that affords us 1M annotated training samples across diverse synthetic scenes. Third, we evaluate different parameterizations for scene flow prediction and adopt a natural and effective parameterization. Our model outperforms existing methods as well as baselines built on large-scale models in terms of 3D end-point error, and shows zero-shot generalization to the casually captured videos from DAVIS and the robotic manipulation scenes from RoboTAP. Overall, our approach makes scene flow prediction more practical in-the-wild.

****

MG-MotionLLM: A Unified Framework for Motion Comprehension and Generation across Multiple Granularities

Bizhu Wu, Jinheng Xie, Keming Shen, Zhe Kong, Jianfeng Ren, Ruibin Bai, Rong Qu, Linlin Shen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27849-27858

Recent motion-aware large language models have demonstrated promising potential in unifying motion comprehension and generation. However, existing approaches primarily focus on coarse-grained motion-text modeling, where text describes the overall semantics of an entire motion sequence in just a few words. This limits their ability to handle fine-grained motion-relevant tasks, such as understanding and controlling the movements of specific body parts. To overcome this limitation, we pioneer MG-MotionLLM, a unified motion-language model for multi-granular motion comprehension and generation. We further introduce a comprehensive multi-granularity training scheme by incorporating a set of novel auxiliary tasks, such as localizing temporal boundaries of motion segments via detailed text as well as motion detailed captioning, to facilitate mutual reinforcement for motion-text modeling across various levels of granularity. Extensive experiments show that our MG-MotionLLM achieves superior performance on classical text-to-motion and motion-to-text tasks, and exhibits potential in novel fine-grained motion comprehension and editing tasks. Project page: CVI-SZU/MG-MotionLLM

****

Retaining Knowledge and Enhancing Long-Text Representations in CLIP through Dual-Teacher Distillation

Yuheng Feng, Changsong Wen, Zelin Peng, Li jiaye, Siyu Zhu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24895-24904

Contrastive language-image pretraining models such as CLIP have demonstrated remarkable performance in various text-image alignment tasks. However, the inherent 77-token input limitation and reliance on predominantly short-text training data restrict its ability to handle long-text tasks effectively. To overcome these constraints, we propose LongD-CLIP, a dual-teacher distillation framework designed to enhance long-text representation while mitigating knowledge forgetting. In our approach, a teacher model, fine-tuned on long-text data, distills rich representation knowledge into a student model, while the original CLIP serves as a s

econdary teacher to help the student retain its foundational knowledge. Extensiv e experiments reveal that LongD-CLIP significantly outperforms existing models a cross long-text retrieval, short-text retrieval, and zero-shot image classificat ion tasks. For instance, in the image-to-text retrieval task on the ShareGPT4V t est set, LongD-CLIP exceeds Long-CLIP's performance by 2.5%, achieving an accura cy of 98.3%. Similarly, on the Urban-1k dataset, it records a 9.2% improvement, reaching 91.9%, thereby underscoring its robust generalization capabilities. Add itionally, the text encoder of LongD-CLIP exhibits reduced latent space drift an d improved compatibility with existing generative models, effectively overcoming the 77-token input constraint.
****

MMRL: Multi-Modal Representation Learning for Vision-Language Models
Yuncheng Guo, Xiaodong Gu; Proceedings of the Computer Vision and Pattern Recogn ition Conference (CVPR), 2025, pp. 25015-25025
Large-scale pre-trained Vision-Language Models (VLMs) have become essential for transfer learning across diverse tasks. However, adapting these models with limi ted few-shot data often leads to overfitting, diminishing their performance on n ew tasks. To tackle this issue, we propose a novel Multi-Modal Representation Le arning (MMRL) framework that introduces a shared, learnable, and modality-agnost ic representation space. MMRL projects the space tokens to text and image repres entation tokens, facilitating more effective multi-modal interactions. Unlike pr evious approaches that solely optimize class token features, MMRL integrates rep resentation tokens at higher layers of the encoders--where dataset-specific feat ures are more prominent--while preserving generalized knowledge in the lower lay ers. During training, both representation and class features are optimized, with trainable projection layer applied to the representation tokens, whereas the cl ass token projection layer remains frozen to retain pre-trained knowledge. Furth ermore, a regularization term is introduced to align the class features and text features with the zero-shot features from the frozen VLM, thereby safeguarding the model's generalization capacity. For inference, a decoupling strategy is emp loyed, wherein both representation and class features are utilized for base clas ses, while only the class features, which retain more generalized knowledge, are used for new tasks. Extensive experiments across 15 datasets demonstrate that M MRL outperforms state-of-the-art methods, achieving a balanced trade-off between task-specific adaptation and generalization. Code is available at https://githu b.com/yunncheng/MMRL.
****

Optical-Flow Guided Prompt Optimization for Coherent Video Generation
Hyelin Nam, Jaemin Kim, Dohun Lee, Jong Chul Ye; Proceedings of the Computer Vis ion and Pattern Recognition Conference (CVPR), 2025, pp. 7837-7846
While text-to-video diffusion models have made significant strides, many still f ace challenges in generating videos with temporal consistency. Within diffusion frameworks, guidance techniques have proven effective in enhancing output qualit y during inference; however, applying these methods to video diffusion models in troduces additional complexity of handling computations across entire sequences. To address this, we propose a novel framework called MotionPrompt that guides t he video generation process via optical flow. Specifically, we train a discrimin ator to distinguish optical flow between random pairs of frames from real videos and generated ones. Given that prompts can influence the entire video, we optim ize learnable token embeddings during reverse sampling steps by using gradients from a trained discriminator applied to random frame pairs. This approach allows our method to generate visually coherent video sequences that closely reflect n atural motion dynamics, without compromising the fidelity of the generated conte nt. We demonstrate the effectiveness of our approach across various models. Proj ect Page: https://motionprompt.github.io
****

MOS: Modeling Object-Scene Associations in Generalized Category Discovery
Zhengyuan Peng, Jinpeng Ma, Zhimin Sun, Ran Yi, Haichuan Song, Xin Tan, Lizhuang Ma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15118-15128

Generalized Category Discovery (GCD) is a classification task that aims to class ify both base and novel classes in unlabeled images, using knowledge from a labe led dataset. In GCD, previous research overlooks scene information or treats it as noise, reducing its impact during model training. However, in this paper, we argue that scene information should be viewed as a strong prior for inferring no vel classes. We attribute the misinterpretation of scene information to a key fa ctor: the Ambiguity Challenge inherent in GCD. Specifically, novel objects in ba se scenes might be wrongly classified into base categories, while base objects i n novel scenes might be mistakenly recognized as novel categories. Once the ambi guity challenge is addressed, scene information can reach its full potential, si gnificantly enhancing the performance of GCD models. To more effectively leverag e scene information, we propose the Modeling Object-Scene Associations (MOS) fra mework, which utilizes a simple MLP-based scene-awareness module to enhance GCD performance. It achieves an exceptional average accuracy improvement of 4% on th e challenging fine-grained datasets compared to state-of-the-art methods, emphas izing its superior performance in fine-grained GCD. The code is publicly availab le at https://github.com/JethroPeng/MOS.

****

Anchor-Aware Similarity Cohesion in Target Frames Enables Predicting Temporal Mo ment Boundaries in 2D

Jiawei Tan, Hongxing Wang, Junwu Weng, Jiaxin Li, Zhilong Ou, Kang Dang; Proceed ings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24180-24189

Video moment retrieval aims to locate specific moments from a video according to the query text. This task presents two main challenges: i) aligning the query a nd video frames at the feature level, and ii) projecting the query-aligned frame features to the start and end boundaries of the matching interval. Previous wor k commonly involves all frames in feature alignment, easy to cause aligning irre levant frames with the query. Furthermore, they forcibly map visual features to interval boundaries but ignoring the information gap between them, yielding subo ptimal performance. In this study, to reduce distraction from irrelevant frames, we designate an anchor frame as that with the maximum query-frame relevance mea sured by the established Vision-Language Model. Via similarity comparison betwee n the anchor frame and the others, we produce a semantically compact segment aro und the anchor frame, which serves as a guide to align features of query and rel ated frames. We observe that such a feature alignment will make similarity cohes ive between target frames, which enables us to predict the interval boundaries b y a single point detection in the 2D semantic similarity space of frames, thus w ell bridging the information gap between frame semantics and temporal boundaries . Experimental results across various datasets demonstrate that our approach sig nificantly improves the alignment between queries and video frames while effecti vely predicting temporal moment boundaries. Especially, on QVHighlights Test and ActivityNet Captions datasets, our proposed approach achieves 3.8% and 7.4% res pectively higher than current state-of-the-art R1@.7 performance. The code is av ailable at https://github.com/ExMorgan-Alter/AFAFSGD.

****

Test-time Augmentation Improves Efficiency in Conformal Prediction

Divya Shanmugam, Helen Lu, Swami Sankaranarayanan, John Guttag; Proceedings of t he Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20622-20 631

A conformal classifier produces a set of predicted classes and provides a probab ilistic guarantee that the set includes the true class. Unfortunately, it is oft en the case that conformal classifiers produce uninformatively large sets. In th is work, we show that test-time augmentation (TTA)---a technique that introduces inductive biases during inference---reduces the size of the sets produced by co nformal classifiers. Our approach is flexible, computationally efficient, and ef fective. It can be combined with any conformal score, requires no model retraini ng, and reduces prediction set sizes by 10%-14% on average. We conduct an evalua tion of the approach spanning three datasets, three models, two established conf ormal scoring methods, different guarantee strengths, and several distribution s

hifts to show when and why test-time augmentation is a useful addition to the co
nformal pipeline.
****

Breaking the Low-Rank Dilemma of Linear Attention
Qihang Fan, Huaibo Huang, Ran He; Proceedings of the Computer Vision and Pattern
 Recognition Conference (CVPR), 2025, pp. 25271-25280
The Softmax attention mechanism in Transformer models is notoriously computation
ally expensive, particularly due to its quadratic complexity, posing significant
 challenges in vision applications. In contrast, linear attention provides a far
 more efficient solution by reducing the complexity to linear levels. However, c
ompared to Softmax attention, linear attention often experiences significant per
formance degradation. Our experiments indicate that this performance drop is due
 to the low-rank nature of linear attention's feature map, which hinders its abi
lity to adequately model complex spatial information. In this paper, to break th
e low-rank dilemma of linear attention, we conduct rank analysis from two perspe
ctives: the kv buffer and the output features. Consequently, we introduce **Rank
-Augmented Linear Attention** (RALA), which rivals the performance of Softmax at
tention while maintaining linear complexity and high efficiency. Based on RALA,
we construct the **Rank-Augmented Vision Linear Transformer** (RAVLT). Extensive
 experiments demonstrate that RAVLT achieves excellent performance across variou
s vision tasks. Specifically, without using any additional labels, data, or supe
rvision during training, RAVLT achieves an **84.4%** Top-1 accuracy on ImageNet-
1k with only **26M** parameters and **4.6G** FLOPs. This result significantly su
rpasses previous linear attention mechanisms, fully illustrating the potential o
f RALA.
****

StoryGPT-V: Large Language Models as Consistent Story Visualizers
Xiaoqian Shen, Mohamed Elhoseiny; Proceedings of the Computer Vision and Pattern
 Recognition Conference (CVPR), 2025, pp. 13273-13283
Recent generative models have demonstrated impressive capabilities in generating
 realistic and visually pleasing images grounded on textual prompts. Neverthele
ss, a significant challenge remains in applying these models for the more intrica
te task of story visualization. Since it requires resolving pronouns (he, she, t
hey) in the frame descriptions, i.e., anaphora resolution, and ensuring consiste
nt characters and background synthesis across frames. Yet, the emerging Large La
nguage Model (LLM) showcases robust reasoning abilities to navigate through ambi
guous references and process extensive sequences. Therefore, we introduce StoryG
PT-V, which leverages the merits of the latent diffusion (LDM) and LLM to produc
e images with consistent and high-quality characters grounded on given story des
criptions. First, we train a character-aware LDM, which takes character-augmente
d semantic embedding as input and includes the supervision of the cross-attentio
n map using character segmentation masks, aiming to enhance character generation
 accuracy and faithfulness. In the second stage, we enable an alignment between
the output of LLM and the character-augmented embedding residing in the input sp
ace of the first-stage model. This harnesses the reasoning ability of LLM to add
ress ambiguous references and the comprehension capability to memorize the conte
xt. We conduct comprehensive experiments on two visual story visualization bench
marks. Our model reports superior quantitative results and consistently generate
s accurate characters of remarkable quality with low memory consumption. Our cod
e is publicly available at: https://xiaoqian-shen.github.io/StoryGPT-V.
****

Code-as-Monitor: Constraint-aware Visual Programming for Reactive and Proactive
Robotic Failure Detection
Enshen Zhou, Qi Su, Cheng Chi, Zhizheng Zhang, Zhongyuan Wang, Tiejun Huang, Lu
Sheng, He Wang; Proceedings of the Computer Vision and Pattern Recognition Confe
rence (CVPR), 2025, pp. 6919-6929
Automatic detection and prevention of open-set failures are crucial in closed-lo
op robotic systems. Recent studies often struggle to simultaneously identify une
xpected failures reactively after they occur and prevent foreseeable ones proact
ively. To this end, we propose Code-as-Monitor (CaM), a novel paradigm leveragin

g the vision-language model (VLM) for both open-set reactive and proactive failu re detection. The core of our method is to formulate both tasks as a unified set of spatio-temporal constraint satisfaction problems and use VLM-generated code to evaluate them for real-time monitoring. To enhance the accuracy and efficienc y of monitoring, we further introduce constraint elements that abstract constrai nt-related entities or their parts into compact geometric elements. This approac h offers greater generality, simplifies tracking, and facilitates constraint-awa re visual programming by leveraging these elements as visual prompts. Experiment s show that CaM achieves a 28.7% higher success rate and reduces execution time by 31.8% under severe disturbances compared to baselines across three simulators and a real-world setting. Moreover, CaM can be integrated with open-loop contro l policies to form closed-loop systems, enabling long-horizon tasks in cluttered scenes with dynamic environments.

****

Embracing Collaboration Over Competition: Condensing Multiple Prompts for Visual In-Context Learning

Jinpeng Wang, Tianci Luo, Yaohua Zha, Yan Feng, Ruisheng Luo, Bin Chen, Tao Dai, Long Chen, Yaowei Wang, Shu-Tao Xia; Proceedings of the Computer Vision and Pat tern Recognition Conference (CVPR), 2025, pp. 25156-25165

Visual In-Context Learning (VICL) enables adaptively solving vision tasks by lev eraging pixel demonstrations, mimicking human-like task completion through analo gy. Prompt selection is critical in VICL, but current methods assume the existen ce of a single "ideal" prompt in a pool of candidates, which in practice may not hold true. Multiple suitable prompts may exist, but individually they often fal l short, leading to difficulties in selection and the exclusion of useful contex t. To address this, we propose a new perspective: prompt condensation. Rather th an relying on a single prompt, candidate prompts collaborate to efficiently inte grate informative contexts without sacrificing resolution. We devise Condenser, a lightweight external plugin that compresses relevant fine-grained context acro ss multiple prompts. Optimized end-to-end with the backbone, Condenser ensures a ccurate integration of contextual cues. Experiments demonstrate Condenser outper forms state-of-the-arts across benchmark tasks, showing superior context compres sion, scalability with more prompts, and enhanced computational efficiency compa red to ensemble methods, positioning it as a highly competitive solution for VIC L. Code is open-sourced at https://github.com/gimpong/CVPR25-Condenser.

****

Rethinking Reconstruction and Denoising in the Dark: New Perspective, General Ar chitecture and Beyond

Tengyu Ma, Long Ma, Ziye Li, Yuetong Wang, Jinyuan Liu, Chengpei Xu, Risheng Liu ; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2323-2332

Recently, enhancing image quality in the original RAW domain has garnered signif icant attention, with denoising and reconstruction emerging as fundamental tasks . Although some works attempt to couple these tasks, they primarily focus on cas cade learning while neglecting task associativity within a broader parameter spa ce, leading to suboptimal performance. This work introduces a novel approach by rethinking denoising and reconstruction from a "backbone-head" perspective, leve raging the stronger shared parameter space offered by the backbone, compared to the encoder used in existing works. We derive task-specific heads with fewer par ameters to mitigate learning pressure. By incorporating chromaticity-and-noise p erception module into the backbone and introducing task-specific supervision dur ing training, we enable simultaneous high-quality results for reconstruction and denoising. Additionally, we design a dual-head interaction module to capture th e latent correspondence between the two tasks, significantly enhancing multi-tas k accuracy. Extensive experiments validate the superiority of the proposed metho d. Code is available at: https://github.com/csmty/CANS.

****

Edge-SD-SR: Low Latency and Parameter Efficient On-device Super-Resolution with Stable Diffusion via Bidirectional Conditioning

Isma Hadji, Mehdi Noroozi, Victor Escorcia, Anestis Zaganidis, Brais Martinez, G

eorgios Tzimiropoulos; Proceedings of the Computer Vision and Pattern Recognitio n Conference (CVPR), 2025, pp. 12789-12798
There has been immense progress recently in the visual quality of Stable Diffusi on-based Super Resolution (SD-SR). However, deploying large diffusion models on computationally restricted devices such as mobile phones remains impractical due to the large model size and high latency. This is compounded for SR as it often operates at high res (e.g. 4Kx3K). In this work, we introduce Edge-SD-SR, the f irst parameter efficient and low latency diffusion model for image super-resolut ion. Edge-SD-SR consists of 169M parameters, including UNet, encoder and decod er, and has a complexity of only 142 GFLOPs. To maintain a high visual quality on such low compute budget, we introduce a number of training strategies: (i) A novel conditioning mechanism on the low-resolution input, coined bidirectional conditioning, which tailors the SD model for the SR task. (ii) Joint training of the UNet and encoder, while decoupling the encodings of the HR and LR images an d using a dedicated schedule. (iii) Finetuning the decoder using the UNet's outp ut to directly tailor the decoder to the latents obtained at inference time. Edg e-SD-SR runs efficiently on device, e.g. it can upscale a 128x128 patch to 512x5 12 in 38 msec while running on a Samsung S24 DSP, and of a 512 x 512 to 2, 048 x 2, 048 (requiring 25 model evaluations) in just 1.1 sec. Furthermore, we show that Edge-SD-SR matches or even outperforms state-of-the-art SR approaches on t he most established SR benchmarks.
****
Unity in Diversity: Video Editing via Gradient-Latent Purification
Junyu Gao, Kunlin Yang, Xuan Yao, Yufan Hu; Proceedings of the Computer Vision a nd Pattern Recognition Conference (CVPR), 2025, pp. 23401-23411
Recently, text-driven video editing methods that optimize target latent represen tations have garnered significant attention and demonstrated promising results. However, these methods rely on self-supervised objectives to compute the gradien ts needed for updating latent representations, which inevitably introduces gradi ent noise, compromising content generation quality. Additionally, it is challeng ing to determine the optimal stopping point for the editing process, making it d ifficult to achieve an optimal solution for the latent representation. To addres s these issues, we propose a unified gradient-latent purification framework that collects gradient and latent information across different stages to identify ef fective and concordant update directions. We design a local coordinate system co nstruction method based on feature decomposition, enabling short-term gradients and final-stage latents to be reprojected onto new axes. Then, we employ tailore d coefficient regularization terms to effectively aggregate the decomposed infor mation. Additionally, a temporal smoothing axis extension strategy is developed to enhance the temporal coherence of the generated content. Extensive experiment s demonstrate that our proposed method outperforms state-of-the-art methods acro ss various editing tasks, delivering superior editing performance. Project page is available in https://unity-in-diversity-editing.github.io.
****
Revealing Key Details to See Differences: A Novel Prototypical Perspective for S keleton-based Action Recognition
Hongda Liu, Yunfan Liu, Min Ren, Hao Wang, Yunlong Wang, Zhenan Sun; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 292 48-29257
In skeleton-based action recognition, a key challenge is distinguishing between actions with similar trajectories of joints due to the lack of image-level detai ls in skeletal representations. Recognizing that the differentiation of similar actions relies on subtle motion details in specific body parts, we direct our ap proach to focus on the fine-grained motion of local skeleton components. To this end, we introduce ProtoGCN, a Graph Convolutional Network (GCN)-based model tha t breaks down the dynamics of entire skeleton sequences into a combination of le arnable prototypes representing core motion patterns of action units. By contras ting the reconstruction of prototypes, ProtoGCN can effectively identify and enh ance the discriminative representation of similar actions. Without bells and whi stles, ProtoGCN achieves state-of-the-art performance on multiple benchmark data

sets, including NTU RGB+D, NTU RGB+D 120, Kinetics-Skeleton, and FineGYM, which demonstrates the effectiveness of the proposed method. The code is available at https://github.com/firework8/ProtoGCN.
****

Finsler Multi-Dimensional Scaling: Manifold Learning for Asymmetric Dimensionality Reduction and Embedding
Thomas Dagès, Simon Weber, Ya-Wei Eileen Lin, Ronen Talmon, Daniel Cremers, Michael Lindenbaum, Alfred M. Bruckstein, Ron Kimmel; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25842-25853
Dimensionality reduction is a fundamental task that aims to simplify complex data by reducing its feature dimensionality while preserving essential patterns, with core applications in data analysis and visualisation. To preserve the underlying data structure, multi-dimensional scaling (MDS) methods focus on preserving pairwise dissimilarities, such as distances. They optimise the embedding to have pairwise distances as close as possible to the data dissimilarities. However, the current standard is limited to embedding data in Riemannian manifolds. Motivated by the lack of asymmetry in the Riemannian metric of the embedding space, this paper extends the MDS problem to a natural asymmetric generalisation of Riemannian manifolds called Finsler manifolds. Inspired by Euclidean space, we define a canonical Finsler space for embedding asymmetric data. Due to its simplicity with respect to geodesics, data representation in this space is both intuitive and simple to analyse. We demonstrate that our generalisation benefits from the same theoretical convergence guarantees. We reveal the effectiveness of our Finsler embedding across various types of non-symmetric data, highlighting its value in applications such as data visualisation, dimensionality reduction, directed graph embedding, and link prediction.
****

VideoEspresso: A Large-Scale Chain-of-Thought Dataset for Fine-Grained Video Reasoning via Core Frame Selection
Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, Si Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26181-26191
The advancement of Large Vision Language Models (LVLMs) has significantly improved multimodal understanding, yet challenges remain in video reasoning tasks due to the scarcity of high-quality, large-scale datasets. Existing video question-answering (VideoQA) datasets often rely on costly manual annotations with insufficient granularity or automatic construction methods with redundant frame-by-frame analysis, limiting their scalability and effectiveness for complex reasoning. To address these challenges, we introduce VideoEspresso, a novel dataset that features VideoQA pairs preserving essential spatial details and temporal coherence, along with multimodal annotations of intermediate reasoning steps. Our construction pipeline employs a semantic-aware method to reduce redundancy, followed by generating QA pairs using GPT-4o. We further develop video Chain-of-Thought (CoT) annotations to enrich reasoning processes, guiding GPT-4o in extracting logical relationships from QA pairs and video content. To exploit the potential of high-quality VideoQA pairs, we propose a Hybrid LVLMs Collaboration framework, featuring a Frame Selector and a two-stage instruction fine-tuned reasoning LVLM. This framework adaptively selects core frames and performs CoT reasoning using multimodal evidence. Evaluated on our proposed benchmark with 14 tasks against 9 popular LVLMs, our method outperforms existing baselines on most tasks, demonstrating superior video reasoning capabilities. Our code and dataset have been released at: https://github.com/hshjerry/VideoEspresso
****

INFP: Audio-Driven Interactive Head Generation in Dyadic Conversations
Yongming Zhu, Longhao Zhang, Zhengkun Rong, Tianshu Hu, Shuang Liang, Zhipeng Ge; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10667-10677
Imagine having a conversation with a socially intelligent agent. It can attentively listen to your words and offer visual and linguistic feedback promptly. This seamless interaction allows for multiple rounds of conversation to flow smoothl

y and naturally. In pursuit of actualizing it, we propose INFP, a novel audio-dr iven head generation framework for dyadic interaction. Unlike previous head gene ration works that only focus on single-sided communication, or require manual ro le assignment and explicit role switching, our model drives the agent portrait d ynamically alternates between speaking and listening state, guided by the input dyadic audio. Specifically, INFP comprises a Motion-Based Head Imitation stage a nd an Audio-Guided Motion Generation stage. The first stage learns to project fa cial communicative behaviors from real-life conversation videos into a low-dimen sional motion latent space, and use the motion latent codes to animate a static image. The second stage learns the mapping from the input dyadic audio to motion latent codes through denoising, leading to the audio-driven head generation in interactive scenarios. To facilitate this line of research, we introduce DyConv, a large scale dataset of rich dyadic conversations collected from the Internet. Extensive experiments and visualizations demonstrate superior performance and e ffectiveness of our method.

****

Federated Learning with Domain Shift Eraser

Zheng Wang, Zihui Wang, Zheng Wang, Xiaoliang Fan, Cheng Wang; Proceedings of th e Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4978-4987

Federated learning (FL) is emerging as a promising technique for collaborative l earning without local data leaving their devices. However, clients' data origina ting from diverse domains may degrade model performance due to domain shifts, pr eventing the model from learning consistent representation space. In this paper, we propose a novel FL framework, Federated Domain Shift Eraser (FDSE), to impro ve model performance by differently erasing each client's domain skew and enhanc ing their consensus. First, we formulate the model forward passing as an iterati ve deskewing process that extracts and then deskews features alternatively. This is efficiently achieved by decomposing each original layer in the neural networ k into a Domain-agnostic Feature Extractor (DFE) and a Domain-specific Skew Eras er (DSE). Then, a regularization term is applied to promise the effectiveness of feature deskewing by pulling local statistics of DSE's outputs close to the glo bally consistent ones. Finally, DFE modules are fairly aggregated and broadcast to all the clients to maximize their consensus, and DSE modules are personalized for each client via similarity-aware aggregation to erase their domain skew dif ferently. Comprehensive experiments were conducted on three datasets to confirm the advantages of our method in terms of accuracy, efficiency, and generalizabil ity.

****

Cross-Modal Distillation for 2D/3D Multi-Object Discovery from 2D Motion

Saad Lahlali, Sandra Kara, Hejer Ammar, Florian Chabot, Nicolas Granger, Hervé L e Borgne, Quoc-Cuong Pham; Proceedings of the Computer Vision and Pattern Recogn ition Conference (CVPR), 2025, pp. 24529-24538

Object discovery, which refers to the task of localizing objects without human a nnotations, has gained significant attention in 2D image analysis. However, desp ite this growing interest, it remains under-explored in 3D data, where approache s rely exclusively on 3D motion, despite its several challenges. In this paper, we present a novel framework that leverages advances in 2D object discovery whic h are based on 2D motion to exploit the advantages of such motion cues being mor e flexible and generalizable and to bridge the gap between 2D and 3D modalities. Our primary contributions are twofold: (i) we introduce DIOD-3D, the first basel ine for multi-object discovery in 3D data using 2D motion, incorporating scene c ompletion as an auxiliary task to enable dense object localization from sparse i nput data; (ii) we develop xMOD, a cross-modal training framework that integrate s 2D and 3D data while always using 2D motion cues. xMOD employs a teacher-stude nt training paradigm across the two modalities to mitigate confirmation bias by leveraging the domain gap. During inference, the model supports both RGB-only an d point cloud-only inputs. Additionally, we propose a late-fusion technique tail ored to our pipeline that further enhances performance when both modalities are available at inference.We evaluate our approach extensively on synthetic (TRIP-P D) and challenging real-world datasets (KITTI and Waymo). Notably, our approach

yields a substantial performance improvement compared with the 2D object discovery state-of-the-art on all datasets with gains ranging from +8.7 to +15.1 in F1@50 score.
****

DiTCtrl: Exploring Attention Control in Multi-Modal Diffusion Transformer for Tuning-Free Multi-Prompt Longer Video Generation
Minghong Cai, Xiaodong Cun, Xiaoyu Li, Wenze Liu, Zhaoyang Zhang, Yong Zhang, Ying Shan, Xiangyu Yue; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7763-7772
Sora-like video generation models have achieved remarkable progress with a Multi-Modal Diffusion Transformer (MM-DiT) architecture. However, the current video generation models predominantly focus on single-prompt, struggling to generate coherent scenes with multiple sequential prompts that better reflect real-world dynamic scenarios. While some pioneering works have explored multi-prompt video generation, they face significant challenges including strict training data requirements, weak prompt following, and unnatural transitions. To address these problems, we propose DiTCtrl, a training-free multi-prompt video generation method under MM-DiT architectures for the first time. Our key idea is to take the multi-prompt video generation task as temporal video editing with smooth transitions. To achieve this goal, we first analyze MM-DiT's attention mechanism, finding that the 3D full attention behaves similarly to that of the cross/self-attention blocks in the UNet-like diffusion models, enabling mask-guided precise semantic control across different prompts with attention sharing for multi-prompt video generation. Based on our careful design, the video generated by DiTCtrl achieves smooth transitions and consistent object motion given multiple sequential prompts without additional training. Besides, we also present MPVBench, a new benchmark specially designed for multi-prompt video generation to evaluate the performance of multi-prompt generation. Extensive experiments demonstrate that our method achieves state-of-the-art performance without additional training.
****

Bridge Frame and Event: Common Spatiotemporal Fusion for High-Dynamic Scene Optical Flow
Hanyu Zhou, Haonan Wang, Haoyue Liu, Yuxing Duan, Yi Chang, Luxin Yan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27904-27913
High-dynamic scene optical flow is a challenging task, which suffers spatial blur and temporal discontinuous motion due to large displacement in frame imaging, thus deteriorating the spatiotemporal feature of optical flow. Typically, existing methods mainly introduce event camera to directly fuse the spatiotemporal features between the two modalities. However, this direct fusion is ineffective, since there exists a large gap due to the heterogeneous data representation between frame and event modalities. To address this issue, we explore a common-latent space as an intermediate bridge to mitigate the modality gap. In this work, we propose a novel common spatiotemporal fusion between frame and event modalities for high-dynamic scene optical flow, including visual boundary localization and motion correlation fusion. Specifically, in visual boundary localization, we figure out that frame and event share the similar spatiotemporal gradients, whose similarity distribution is consistent with the extracted boundary distribution. This motivates us to design the common spatiotemporal gradient to constrain the reference boundary localization. In motion correlation fusion, we discover that the frame-based motion possesses spatially dense but temporally discontinuous correlation, while the event-based motion has spatially sparse but temporally continuous correlation. This inspires us to use the reference boundary to guide the complementary motion knowledge fusion between the two modalities. Moreover, common spatiotemporal fusion can not only relieve the cross-modal feature discrepancy, but also make the fusion process interpretable for dense and continuous optical flow. Extensive experiments have been performed to verify the superiority of the proposed method.
****

EVPGS: Enhanced View Prior Guidance for Splatting-based Extrapolated View Synthe

sis
Jiahe Li, Feiyu Wang, Xiaochao Qu, Chengjing Wu, Luoqi Liu, Ting Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16398-16407
Gaussian Splatting (GS)-based methods rely on sufficient training view coverage and perform synthesis on interpolated views. In this work, we tackle the more challenging and underexplored Extrapolated View Synthesis (EVS) task. Here we enable GS-based models trained with limited view coverage to generalize well to extrapolated views. To achieve our goal, we propose a view augmentation framework to guide training through a coarse-to-fine process. At the coarse stage, we reduce rendering artifacts due to insufficient view coverage by introducing a regularization strategy at both appearance and geometry levels. At the fine stage, we generate reliable view priors to provide further training guidance. To this end, we incorporate an occlusion awareness into the view prior generation process, and refine the view priors with the aid of coarse stage output. We call our framework Enhanced View Prior Guidance for Splatting (EVPGS). To comprehensively evaluate EVPGS on the EVS task, we collect a real-world dataset called Merchandise3D dedicated to the EVS scenario. Experiments on three datasets including both real and synthetic demonstrate EVPGS achieves state-of-the-art performance, while improving synthesis quality at extrapolated views for GS-based methods both qualitatively and quantitatively. We will make our code, dataset, and models public.
****
GREAT: Geometry-Intention Collaborative Inference for Open-Vocabulary 3D Object Affordance Grounding
Yawen Shao, Wei Zhai, Yuhang Yang, Hongchen Luo, Yang Cao, Zheng-Jun Zha; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17326-17336
Open-Vocabulary 3D object affordance grounding aims to anticipate "action possibilities" regions on 3D objects with arbitrary instructions, which is crucial for robots to generically perceive real scenarios and respond to operational changes. Existing methods focus on combining images or languages that depict interactions with 3D geometries to introduce external interaction priors. However, they are still vulnerable to a limited semantic space by failing to leverage implied invariant geometries and potential interaction intentions. Normally, humans address complex tasks through multi-step reasoning and respond to diverse situations by leveraging associative and analogical thinking. In light of this, we propose GREAT (Geometry-Intention Collaborative Inference) for Open-Vocabulary 3D Object Affordance Grounding, a novel framework that mines the object invariant geometry attributes and performs analogically reason in potential interaction scenarios to form affordance knowledge, fully combining the knowledge with both geometries and visual contents to ground 3D object affordance. Besides, we introduce the Point Image Affordance Dataset v2 (PIADv2), the largest 3D object affordance dataset at present to support the task. Extensive experiments demonstrate the effectiveness and superiority of GREAT. The code and dataset are available at https://yawen-shao.github.io/GREAT/.
****
Link to the Past: Temporal Propagation for Fast 3D Human Reconstruction from Monocular Video
Matthew Marchellus, Nadhira Noor, In Kyu Park; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6190-6199
Fast 3D clothed human reconstruction from monocular video remains a significant challenge in computer vision, particularly in balancing computational efficiency with reconstruction quality. Current approaches are either focused on static image reconstruction but too computationally intensive, or achieve high quality through per-video optimization that requires minutes to hours of processing, making them unsuitable for real-time applications. To this end, we present TemPoFast3D, a novel method that leverages temporal coherency of human appearance to reduce redundant computation while maintaining reconstruction quality. Our approach is a "plug-and play" solution that uniquely transforms pixel-aligned reconstruction networks to handle continuous video streams by maintaining and refining a can

onical appearance representation through efficient coordinate mapping. Extensive experiments demonstrate that TemPoFast3D matches or exceeds state-of-the-art methods across standard metrics while providing high-quality textured reconstruction across diverse pose and appearance, with a maximum speed of 12 FPS.
****

Inversion Circle Interpolation: Diffusion-based Image Augmentation for Data-scarce Classification
Yanghao Wang, Long Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25560-25569
Data Augmentation (DA), i.e., synthesizing faithful and diverse samples to expand the original training set, is a prevalent and effective strategy to improve the performance of various data-scarce tasks. With the powerful image generation ability, diffusion-based DA has shown strong performance gains on different image classification benchmarks. In this paper, we analyze today's diffusion-based DA methods, and argue that they cannot take account of both faithfulness and diversity, which are two critical keys for generating high-quality samples and boosting classification performance. To this end, we propose a novel Diffusion-based DA method: Diff-II. Specifically, it consists of three steps: 1) Category concepts learning: Learning concept embeddings for each category. 2) Inversion interpolation: Calculating the inversion for each image, and conducting circle interpolation for two randomly sampled inversions from the same category. 3) Two-stage denoising: Using different prompts to generate synthesized images in a coarse-to-fine manner. Extensive experiments on various data-scarce image classification tasks (e.g., few-shot, long-tailed, and out-of-distribution classification) have demonstrated its effectiveness over state-of-the-art diffusion-based DA methods.
****

Deterministic Certification of Graph Neural Networks against Graph Poisoning Attacks with Arbitrary Perturbations
Jiate Li, Meng Pang, Yun Dong, Binghui Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5020-5029
Graph neural networks (GNNs) are becoming the de facto method to learn on the graph data and have achieved the state-of-the-art on node and graph classification tasks. However, recent works show GNNs are vulnerable to training-time poisoning attacks -- marginally perturbing edges, nodes, and node features of training graphs can largely degrade the GNN performance. Most previous defenses against such attacks are empirical and are soon broken by adaptive / stronger ones. A few provable defenses provide robustness guarantees, but have large gaps when applied in practice: 1) all restrict the attacker's capability to only one type of perturbation; 2) all are designed for a particular GNN task; and 3) their robustness guarantees are not 100% accurate. In this work, we bridge all these gaps by developing PGNNCert, the first certified defense of GNNs against poisoning attacks under arbitrary (edge, node, and node feature) perturbations with deterministic (100% accurate) robustness guarantees. PGNNCert is also applicable to the two most widely-studied node and graph classification tasks. Extensive evaluations on multiple node and graph classification datasets and GNNs demonstrate the effectiveness of PGNNCert to provably defend against arbitrary poisoning perturbations. PGNNCert is also shown to significantly outperform the state-of-the-art certified defenses against edge perturbation or node perturbation during GNN training.
****

A3: Few-shot Prompt Learning of Unlearnable Examples with Cross-Modal Adversarial Feature Alignment
Xuan Wang, Xitong Gao, Dongping Liao, Tianrui Qin, Yu-liang Lu, Cheng-zhong Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9507-9516
In the age of pervasive machine learning applications, protecting digital content from unauthorized use has become a pressing concern. Unlearnable examples (UEs)--data modified with imperceptible perturbations to inhibit model training while preserving human usability--have emerged as a promising approach. However, existing UE methods assume unauthorized trainers have extensive exposure to UEs or

that models are trained from scratch, which may not hold in practical scenarios. This paper investigates the effectiveness of UEs under the few-shot learning paradigm, pitching it against prompt learning (PL) models that leverage pretrained vision-language models (VLMs), like CLIP, capable of generalizing to new classes with minimal data. To address this, we introduce an adaptive UE framework to generate unlearnable examples that specifically target the PL process. In addition, we propose a novel UE countermeasure, A3, with cross-modal adversarial feature alignment, specifically designed to circumvent UEs under few-shot PL. Experimental evaluations on 7 datasets show that A3 outperforms existing PL methods, achieving up to 33% higher performance in learning from UEs. For example, in the scenario involving l_infinity-bounded EM perturbations, A3 has an average harmonic mean accuracy across 7 datasets of 82.43%, compared to CoCoOp's baseline of 65.47%. Our findings highlight the limitations of existing UEs against PL and lay the foundation for future data protection mechanisms.

****

Adapting Pre-trained 3D Models for Point Cloud Video Understanding via Cross-frame Spatio-temporal Perception

Baixuan Lv, Yaohua Zha, Tao Dai, Xue Yuerong, Ke Chen, Shu-Tao Xia; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12413-12422

Point cloud video understanding is becoming increasingly important in fields such as robotics, autonomous driving, and augmented reality, as they can accurately represent object motion and environmental changes. Despite the progress made in self-supervised learning methods for point cloud video understanding, the limited availability of 4D data and the high computational cost of training 4D-specific models remain significant obstacles. In this paper, we investigate the potential of transferring pre-trained static 3D point cloud models to the 4D domain, identifying the limitations of static models that capture only spatial information while neglecting temporal dynamics. To address this, we propose a novel Cross-frame Spatio-temporal Adaptation (CSA) strategy by introducing the Point Tube Adapter as the embedding layer and the Geometric Constraint Temporal Adapter (GCTA) to enforce temporal consistency across frames. This strategy extracts both short-term and long-term temporal dynamics, effectively integrating them with spatial features and enriching the model's understanding of temporal changes in point cloud videos. Extensive experiments on 3D action and gesture recognition tasks demonstrate that our method achieves state-of-the-art performance, establishing its effectiveness for point cloud video understanding.

****

MASH-VLM: Mitigating Action-Scene Hallucination in Video-LLMs through Disentangled Spatial-Temporal Representations

Kyungho Bae, Jinhyung Kim, Sihaeng Lee, Soonyoung Lee, Gunhee Lee, Jinwoo Choi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13744-13753

In this work, we tackle action-scene hallucination in Video Large Language Models (Video-LLMs), where models incorrectly predict actions based on the scene context or scenes based on observed actions. We observe that existing Video-LLMs often suffer from action-scene hallucination due to two main factors. First, existing Video-LLMs intermingle spatial and temporal features by applying an attention operation across all tokens. Second, they use the standard Rotary Position Embedding (RoPE), which causes the text tokens to overemphasize certain types of tokens depending on their sequential orders. To address these issues, we introduce MASH-VLM, Mitigating Action-Scene Hallucination in Video-LLMs through disentangled spatial-temporal representations. Our approach includes two key innovations: (1) DST-attention, a novel attention mechanism that disentangles the spatial and temporal tokens within the LLM by using masked attention to restrict direct interactions between the spatial and temporal tokens; (2) Harmonic-RoPE, which extends the dimensionality of the positional IDs, allowing the spatial and temporal tokens to maintain balanced positions relative to the text tokens. To evaluate the action-scene hallucination in Video-LLMs, we introduce the UNSCENE benchmark with 1,320 videos and 4,078 QA pairs. Extensive experiments demonstrate that MAS

H-VLM achieves state-of-the-art results on the UNSCENE benchmark, as well as on existing video understanding benchmarks.
****

UWAV: Uncertainty-weighted Weakly-supervised Audio-Visual Video Parsing
Yung-Hsuan Lai, Janek Ebbers, Yu-Chiang Frank Wang, François Germain, Michael Jeffrey Jones, Moitreya Chatterjee; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13561-13570
Audio-Visual Video Parsing (AVVP) entails the challenging task of localizing both uni-modal events (i.e., those occurring exclusively in either the visual or acoustic modality of a video) and multi-modal events (i.e., those occurring in both modalities concurrently). Moreover, the prohibitive cost of annotating training data with the class labels of all these events, along with their start and end times, imposes constraints on the scalability of AVVP techniques unless they can be trained in a weakly-supervised setting, where only modality-agnostic, video-level labels are available in the training data. To this end, recently proposed approaches seek to generate segment-level pseudo-labels to better guide model training. However, the absence of inter-segment dependencies when generating these pseudo-labels and the general bias towards predicting labels that are absent in a segment limit their performance. This work proposes a novel approach towards overcoming these weaknesses called Uncertainty-weighted Weakly-supervised Audio-visual Video Parsing (UWAV). Additionally, our innovative approach factors in the uncertainty associated with these estimated pseudo-labels and incorporates a feature mixup based training regularization for improved training. Empirical results show that UWAV outperforms state-of-the-art methods for the AVVP task on multiple metrics, across two different datasets, attesting to its effectiveness and generalizability.
****

Mosaic of Modalities: A Comprehensive Benchmark for Multimodal Graph Learning
Jing Zhu, Yuhang Zhou, Shengyi Qian, Zhongmou He, Tong Zhao, Neil Shah, Danai Koutra; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14215-14224
Graph machine learning has made significant strides in recent years, yet the integration of visual information with graph structure and its potential for improving performance in downstream tasks remains an underexplored area. To address this critical gap, we introduce the Multimodal Graph Benchmark (MM-GRAPH), a pioneering benchmark that incorporates both visual and textual information into graph learning tasks. MM-GRAPH extends beyond existing text-attributed graph benchmarks, offering a more comprehensive evaluation framework for multimodal graph learning. Our benchmark comprises seven diverse datasets of varying scales (ranging from thousands to millions of edges), designed to assess algorithms across different tasks in real-world scenarios. These datasets feature rich multimodal node attributes, including visual data, which enables a more holistic evaluation of various graph learning frameworks in complex, multimodal environments. To support advancements in this emerging field, we provide an extensive empirical study on various graph learning frameworks when presented with features from multiple modalities, particularly emphasizing the impact of visual information. This study offers valuable insights into the challenges and opportunities of integrating visual data into graph learning.
****

TSP-Mamba: The Travelling Salesman Problem Meets Mamba for Image Super-resolution and Beyond
Kun Zhou, Xinyu Lin, Jiangbo Lu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28134-28143
Recently, Mamba-based frameworks have achieved substantial advancements across diverse computer vision and NLP tasks, particularly in their capacity for reasoning over long-range information with linear complexity. However, the fixed 2D-to-1D scanning pattern overlooks the local structures of an image, limiting its effectiveness in aggregating 2D spatial information. While stacking additional Mamba layers can partially address this issue, it increases parameter intensity and constrains real-time application. In this work, we reconsider the local optimal

scanning path in Mamba, enhancing the rigid and uniform 1D scan through the loca l shortest path theory, thus creating a structure-aware Mamba suited for lightwe ight single-image super-resolution. Specifically, we draw inspiration from the T raveling Salesman Problem (TSP) to establish a local optimal scanning path for i mproved structural 2D information utilization. Here, local patch aggregation occ urs in a content-adaptive manner with minimal propagation cost. TSP-Mamba demons trates substantial improvements over existing Mamba-based and Transformer-based architectures. For example, TSP-Mamba surpasses MambaIR by up to 0.7dB in lightw eight SISR, with comparable parameters and very slightly extra computational dem ands (1-2 GFlops for 720P images).
****

MVPaint: Synchronized Multi-View Diffusion for Painting Anything 3D
Wei Cheng, Juncheng Mu, Xianfang Zeng, Xin Chen, Anqi Pang, Chi Zhang, Zhibin Wa ng, Bin Fu, Gang Yu, Ziwei Liu, Liang Pan; Proceedings of the Computer Vision an d Pattern Recognition Conference (CVPR), 2025, pp. 585-594
Texturing is a crucial step in the 3D asset production workflow, which enhances the visual appeal and diversity of 3D assets. Despite recent advancements in Tex t-to-Texture (T2T) generation, existing methods often yield subpar results, prim arily due to local discontinuities, inconsistencies across multiple views, and t heir heavy dependence on UV unwrapping outcomes. To tackle these challenges, we propose a novel generation-refinement 3D texturing framework called MVPaint, whi ch can generate high-resolution, seamless textures while emphasizing multi-view consistency. MVPaint mainly consists of three key modules. 1) Synchronized Multi -view Generation (SMG). Given a 3D mesh model, MVPaint first simultaneously gene rates multi-view images by employing an SMG model, which leads to coarse texturi ng results with unpainted parts due to missing observations. 2) Spatial-aware 3D Inpainting (S3I). To ensure complete 3D texturing, we introduce the S3I method, specifically designed to effectively texture previously unobserved areas. 3) UV Refinement (UVR). Furthermore, MVPaint employs a UVR module to improve the text ure quality in the UV space, which first performs a UV-space Super-Resolution, f ollowed by a Spatial-aware Seam-Smoothing algorithm for revising spatial texturi ng discontinuities caused by UV unwrapping. Moreover, we establish two T2T evalu ation benchmarks: the Objaverse T2T benchmark and the GSO T2T benchmark, based o n selected high-quality 3D meshes from the Objaverse dataset and the entire GSO dataset, respectively. Extensive experimental results demonstrate that MVPaint s urpasses existing state-of-the-art methods. Notably, MVPaint could generate high -fidelity textures with minimal Janus issues and highly enhanced cross-view cons istency.
****

FirePlace: Geometric Refinements of LLM Common Sense Reasoning for 3D Object Pla cement
Ian Huang, Yanan Bao, Karen Truong, Howard Zhou, Cordelia Schmid, Leonidas Guiba s, Alireza Fathi; Proceedings of the Computer Vision and Pattern Recognition Con ference (CVPR), 2025, pp. 13466-13476
Scene generation with 3D assets presents a complex challenge, requiring both hig h-level semantic understanding and low-level geometric reasoning. While Multimod al Large Language Models (MLLMs) excel at semantic tasks, their application to 3 D scene generation is hindered by their limited grounding on 3D geometry. In thi s paper, we investigate how to best work with MLLMs in an object placement task. Towards this goal, we introduce a novel framework, FirePlace, that applies exis ting MLLMs in (1) 3D geometric reasoning and the extraction of relevant geometri c details from the 3D scene, (2) constructing and solving geometric constraints on the extracted low-level geometry, and (3) pruning for final placements that c onform to common sense. By combining geometric reasoning with real-world underst anding of MLLMs, our method can propose object placements that satisfy both geom etric constraints as well as high-level semantic common-sense considerations. Ou r experiments show that these capabilities allow our method to place objects mor e effectively in complex scenes with intricate geometry, surpassing the quality of prior work.
****

RENO: Real-Time Neural Compression for 3D LiDAR Point Clouds
Kang You, Tong Chen, Dandan Ding, M. Salman Asif, Zhan Ma; Proceedings of the Co mputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22172-22181
Despite the substantial advancements demonstrated by learning-based neural model s in the LiDAR Point Cloud Compression (LPCC) task, realizing real-time compress ion--an indispensable criterion for numerous industrial applications--remains a formidable challenge. This paper proposes RENO, the first real-time neural codec for 3D LiDAR point clouds, achieving superior performance with a lightweight mo del. RENO skips the octree construction and directly builds upon the multiscale sparse tensor representation. Instead of the multi-stage inferring, RENO devises sparse occupancy codes, which exploit cross-scale correlation and derive voxels ' occupancy in a one-shot manner, greatly saving processing time. Experimental r esults demonstrate that the proposed RENO achieves real-time coding speed, 10 fp s at 14-bit depth on a desktop platform (e.g., one RTX 3090 GPU) for both encodi ng and decoding processes, while providing 12.25% and 48.34% bit-rate savings co mpared to G-PCCv23 and Draco, respectively, at a similar quality. RENO model siz e is merely 1MB, making it attractive for practical applications. The source cod e is available at https://github.com/NJUVISION/RENO.
****

End-to-End Implicit Neural Representations for Classification
Alexander Gielisse, Jan van Gemert; Proceedings of the Computer Vision and Patte rn Recognition Conference (CVPR), 2025, pp. 18728-18737
Implicit neural representations (INRs) such as NeRF and SIREN encode a signal i n neural network parameters and show excellent results for signal reconstruction . Using INRs for downstream tasks, such as classification, is however not straig htforward. Inherent symmetries in the parameters pose challenges and current wor ks primarily focus on designing architectures that are equivariant to these symm etries. However, INR-based classification still significantly under-performs com pared to pixel-based methods like CNNs. This work presents an end-to-end strateg y for initializing SIRENs together with a learned learning-rate scheme, to yield representations that improve classification accuracy. We show that a simple, st raightforward, Transformer model applied to a meta-learned SIREN, without incorp orating explicit symmetry equivariances, outperforms the current state-of-the-ar t. On the CIFAR-10 SIREN classification task, we improve the state-of-the-art wi thout augmentations from 38.8% to 59.6%, and from 63.4% to 64.7% with augmentati ons. We demonstrate scalability on the high-resolution Imagenette dataset achiev ing reasonable reconstruction quality with a classification accuracy of 60.8% an d are the first to do INR classification on the full ImageNet-1K dataset where w e achieve a SIREN classification performance of 23.6%. To the best of our knowle dge, no other SIREN classification approach has managed to set a classification baseline for any high-resolution image dataset. Our code is available at https:/ /github.com/SanderGielisse/MWT
****

ASAP: Advancing Semantic Alignment Promotes Multi-Modal Manipulation Detecting a nd Grounding
Zhenxing Zhang, Yaxiong Wang, Lechao Cheng, Zhun Zhong, Dan Guo, Meng Wang; Proc eedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4005-4014
We present ASAP, a new framework for detecting and grounding multi-modal media m anipulation (DGM4).Upon thorough examination, we observe that accurate fine-grai ned cross-modal semantic alignment between the image and text is vital for accur ately manipulation detection and grounding. While existing DGM4 methods pay rare attention to the cross-modal alignment, hampering the accuracy of manipulation detecting to step further. To remedy this issue, this work targets to advance th e semantic alignment learning to promote this task. Particularly, we utilize the off-the-shelf Multimodal Large-Language Models (MLLMs) and Large Language Model s (LLMs) to construct paired image-text pairs, especially for the manipulated in stances. Subsequently, a cross-modal alignment learning is performed to enhance the semantic alignment. Besides the explicit auxiliary clues, we further design a Manipulation-Guided Cross Attention (MGCA) to provide implicit guidance for au

gmenting the manipulation perceiving. With the grounding truth available during training, MGCA encourages the model to concentrate more on manipulated components while downplaying normal ones, enhancing the model's ability to capture manipulations. Extensive experiments are conducted on the DGM4 dataset, the results demonstrate that our model can surpass the comparison method with a clear margin.
****

UNICL-SAM: Uncertainty-Driven In-Context Segmentation with Part Prototype Discovery

Dianmo Sheng, Dongdong Chen, Zhentao Tan, Qiankun Liu, Qi Chu, Tao Gong, Bin Liu, Jing Han, Wenbin Tu, Shengwei Xu, Nenghai Yu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20201-20211

Recent advancements in in-context segmentation generalists have demonstrated significant success in performing various image segmentation tasks using a limited number of labeled example images. However, real-world applications present challenges due to the variability of support examples, which often exhibit quality issues resulting from various sources and inaccurate labeling. How to extract more robust representations from these examples has always been one of the goals of in-context visual learning. In response, we propose UNICL-SAM, to better model the example distribution and extract robust representations to help in-context segmentation. We incorporate an uncertainty probabilistic module to quantify each example's reliability during both the training and testing phases. Utilizing this uncertainty estimation, we introduce an uncertainty-guided graph augmentation and feature refinement strategy, aimed at mitigating the impact of high-uncertainty regions to enhance the learning of robust representations. Subsequently, we construct prototypes for each example by aggregating part information, thereby creating reliable in-context instruction that effectively represents fine-grained local semantics. This approach serves as a valuable complement to traditional global pooling features. Experimental results demonstrate the effectiveness of the proposed framework, underscoring its potential for real-world applications.
****

Layered Motion Fusion: Lifting Motion Segmentation to 3D in Egocentric Videos

Vadim Tschernezki, Diane Larlus, Iro Laina, Andrea Vedaldi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17637-17648

Computer vision is largely based on 2D techniques, with 3D vision still relegated to a relatively narrow subset of applications. However, by building on recent advances in 3D models such as neural radiance fields, some authors have shown that 3D techniques can at last improve outputs extracted from independent 2D views, by fusing them into 3D and denoising them. This is particularly helpful in egocentric videos, where the camera motion is significant, but only under the assumption that the scene itself is static. In fact, as shown in the recent analysis conducted by EPIC Fields, 3D techniques are ineffective when it comes to studying dynamic phenomena, and, in particular, when segmenting moving objects. In this paper, we look into this issue in more detail. First, we propose to improve dynamic segmentation in 3D by fusing motion segmentation predictions from a 2D-based model into layered radiance fields (Layered Motion Fusion). However, the high complexity of long, dynamic videos makes it challenging to capture the underlying geometric structure, and, as a result, hinders the fusion of motion cues into the (incomplete) scene geometry. We address this issue through test-time refinement, which helps the model to focus on specific frames, thereby reducing the data complexity. This results in a synergy between motion fusion and the refinement, and in turn leads to segmentation predictions of the 3D model that surpass the 2D baseline by a large margin. This demonstrates that 3D techniques can enhance 2D analysis even for dynamic phenomena in a challenging and realistic setting.
****

FADE: Frequency-Aware Diffusion Model Factorization for Video Editing

Yixuan Zhu, Haolin Wang, Shilin Ma, Wenliang Zhao, Yansong Tang, Lei Chen, Jie Zhou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28426-28435

Recent advancements in diffusion frameworks have significantly enhanced video editing, achieving high fidelity and strong alignment with textual prompts. Howeve

r, conventional approaches using image diffusion models fall short in handling v ideo dynamics, particularly for challenging temporal edits like motion adjustmen ts. While current video diffusion models produce high-quality results, adapting them for efficient editing remains difficult due to the heavy computational dema nds that prevent the direct application of previous image editing techniques. To overcome these limitations, we introduce FADE--a training-free yet highly effec tive video editing approach that fully leverages the inherent priors from pre-tr ained video diffusion models via frequency-aware factorization. Rather than simp ly using these models, we first analyze the attention patterns within the video model to reveal how video priors are distributed across different components. Bu ilding on these insights, we propose a factorization strategy to optimize each c omponent's specialized role. Furthermore, we devise spectrum-guided modulation t o refine the sampling trajectory with frequency domain cues, preventing informat ion leakage and supporting efficient, versatile edits while preserving the basic spatial and temporal structure. Extensive experiments on real-world videos demo nstrate that our method consistently delivers high-quality, realistic and tempor ally coherent editing results both qualitatively and quantitatively. Code is ava ilable at https://github.com/EternalEvan/FADE.

****

MotiF: Making Text Count in Image Animation with Motion Focal Loss

Shijie Wang, Samaneh Azadi, Rohit Girdhar, Saketh Rambhatla, Chen Sun, Xi Yin; P roceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 202 5, pp. 7773-7783

Text-Image-to-Video (TI2V) generation aims to generate a video from an image fol lowing a text description, which is also referred to as text-guided image animat ion. Most existing methods struggle to generate videos that align well with the text prompts, particularly when motion is specified. To overcome this limitation , we introduce MotiF, a simple yet effective approach that directs the model's l earning to the regions with more motion, thereby improving the text alignment an d motion generation. We use optical flow to generate a motion heatmap and weight the loss according to the intensity of the motion. This modified objective lead s to noticeable improvements and complements existing methods that utilize motio n priors as model inputs. Additionally, due to the lack of a diverse benchmark for evaluating TI2V generation, we propose TI2V-Bench, a dataset consists of 320 image-text pairs for robust evaluation. We present a human evaluation protocol that asks the annotators to select an overall preference between two videos foll owed by their justifications. Through a comprehensive evaluation, MotiF outperfo rms nine open-sourced models, achieving an average preference of 72%. The TI2V-B ench and additional results are released in https://wang-sj16.github.io/motif/.

****

Towards Explicit Geometry-Reflectance Collaboration for Generalized LiDAR Segmen tation in Adverse Weather

Longyu Yang, Ping Hu, Shangbo Yuan, Lu Zhang, Jun Liu, Hengtao Shen, Xiaofeng Zh u; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 139-149

Existing LiDAR semantic segmentation models often suffer from decreased accuracy when exposed to adverse weather conditions. Recent methods addressing this issu e focus on enhancing training data through weather simulation or universal augme ntation techniques. However, few works have studied the negative impacts caused by the heterogeneous domain shifts in the geometric structure and reflectance in tensity of point clouds. In this paper, we delve into this challenge and address it with a novel Geometry-Reflectance Collaboration (GRC) framework that explici tly separates feature extraction for geometry and reflectance. Specifically, GRC employs a dual-branch architecture designed to process geometric and reflectanc e features independently initially, thereby capitalizing on their distinct chara cteristic. Then, GRC adopts a robust multi-level feature collaboration module to suppress redundant and unreliable information from both branches. Consequently, without complex simulation or augmentation, our method effectively extracts int rinsic information about the scene while suppressing interference, thus achievin g better robustness and generalization in adverse weather conditions. We demonst

rate the effectiveness of GRC through comprehensive experiments on challenging b
enchmarks, showing that our method outperforms previous approaches and establish
es new state-of-the-art results.
****

Data Synthesis with Diverse Styles for Face Recognition via 3DMM-Guided Diffusio
n

Yuxi Mi, Zhizhou Zhong, Yuge Huang, Qiuyang Yuan, Xuan Zhao, Jianqing Xu, Shouho
ng Ding, Shaoming Wang, Rizen Guo, Shuigeng Zhou; Proceedings of the Computer Vi
sion and Pattern Recognition Conference (CVPR), 2025, pp. 21203-21214

Identity-preserving face synthesis aims to generate synthetic face images of vir
tual subjects that can substitute real-world data for training face recognition
models. While prior arts strive to create images with consistent identities and
diverse styles, they face a trade-off between them. Identifying their limitation
 of treating style variation as subject-agnostic and observing that real-world p
ersons actually have distinct, subject-specific styles, this paper introduces Mo
rphFace, a diffusion-based face generator. The generator learns fine-grained fac
ial styles, e.g., shape, pose and expression, from the renderings of a 3D morpha
ble model (3DMM). It also learns identities from an off-the-shelf recognition mo
del. To create virtual faces, the generator is conditioned on novel identities o
f unlabeled synthetic faces, and novel styles that are statistically sampled fro
m a real-world prior distribution. The sampling especially accounts for both int
ra-subject variation and subject distinctiveness. A context blending strategy is
 employed to enhance the generator's responsiveness to identity and style condit
ions. Extensive experiments show that MorphFace outperforms the best prior arts
in face recognition efficacy.
****

Diffusion Self-Distillation for Zero-Shot Customized Image Generation

Shengqu Cai, Eric Ryan Chan, Yunzhi Zhang, Leonidas Guibas, Jiajun Wu, Gordon We
tzstein; Proceedings of the Computer Vision and Pattern Recognition Conference (
CVPR), 2025, pp. 18434-18443

Text-to-image diffusion models produce impressive results but are frustrating to
ols for artists who desire fine-grained control. For example, a common use case
is to create images of a specific instance in novel contexts, i.e., "identity-pr
eserving generation". This setting, along with many other tasks (e.g., relightin
g), is a natural fit for image+text-conditional generative models. However, ther
e is insufficient high-quality paired data to train such a model directly. We pr
opose Diffusion Self-Distillation, a method for using a pre-trained text-to-imag
e model to generate its own dataset for text-conditioned image-to-image tasks. W
e first leverage a text-to-image diffusion model's in-context generation ability
 to create grids of images and curate a large paired dataset with the help of a
Visual-Language Model. We then fine-tune the text-to-image model into a text+ima
ge-to-image model using the curated paired dataset. We demonstrate that Diffusio
n Self-Distillation outperforms existing zero-shot methods and is competitive wi
th per-instance tuning techniques on a wide range of identity-preservation gener
ation tasks, without requiring test-time optimization.
****

Uncertainty-guided Perturbation for Image Super-Resolution Diffusion Model

Leheng Zhang, Weiyi You, Kexuan Shi, Shuhang Gu; Proceedings of the Computer Vis
ion and Pattern Recognition Conference (CVPR), 2025, pp. 17980-17989

Diffusion-based image super-resolution methods have demonstrated significant adv
antages over GAN-based approaches, particularly in terms of perceptual quality.
Building upon a lengthy Markov chain, diffusion-based methods possess remarkable
 modeling capacity, enabling them to achieve outstanding performance in real-wor
ld scenarios. Unlike previous methods that focus on modifying the noise schedule
 or sampling process to enhance performance, our approach emphasizes the improve
d utilization of LR information. We find that different regions of the LR image
can be viewed as corresponding to different timesteps in a diffusion process, wh
ere flat areas are closer to the target HR distribution but edge and texture reg
ions are farther away. In these flat areas, applying a slight noise is more adva
ntageous for the reconstruction. We associate this characteristic with uncertain

ty and propose to apply uncertainty estimate to guide region-specific noise level control, a technique we refer to as Uncertainty-guided Noise Weighting. Pixels with lower uncertainty (i.e., flat regions) receive reduced noise to preserve more LR information, therefore improving performance. Furthermore, we modify the network architecture of previous methods to develop our Uncertainty-guided Perturbation Super-Resolution (UPSR) model. Extensive experimental results demonstrate that, despite reduced model size and training overhead, the proposed UPSR method outperforms current state-of-the-art methods across various datasets, both quantitatively and qualitatively.
****

Geometric Knowledge-Guided Localized Global Distribution Alignment for Federated Learning
Yanbiao Ma, Wei Dai, Wenke Huang, Jiayi Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20958-20968
Data heterogeneity in federated learning, characterized by a significant misalignment between local and global distributions, leads to divergent local optimization directions and hinders global model training. Existing studies mainly focus on optimizing local updates or global aggregation, but these indirect approaches demonstrate instability when handling highly heterogeneous data distributions, especially in scenarios where label skew and domain skew coexist. To address this, we propose a geometry-guided data generation method that centers on simulating the global embedding distribution locally. We first introduce the concept of the geometric shape of an embedding distribution and then address the challenge of obtaining global geometric shapes under privacy constraints. Subsequently, we propose GGEUR, which leverages global geometric shapes to guide the generation of new samples, enabling a closer approximation to the ideal global distribution. In single-domain scenarios, we augment samples based on global geometric shapes to enhance model generalization; in multi-domain scenarios, we further employ class prototypes to simulate the global distribution across domains. Extensive experimental results demonstrate that our method significantly enhances the performance of existing approaches in handling highly heterogeneous data, including scenarios with label skew, domain skew, and their coexistence.
****

Towards Human-Understandable Multi-Dimensional Concept Discovery
Arne Grobrügge, Niklas Kühl, Gerhard Satzger, Philipp Spitzer; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20018-20027
Concept-based eXplainable AI (C-XAI) aims to overcome the limitations of traditional saliency maps by converting pixels into human-understandable concepts that are consistent across an entire dataset. A crucial aspect of C-XAI is completeness, which measures how well a set of concepts explains a model's decisions. Among C-XAI methods, Multi-Dimensional Concept Discovery (MCD) effectively improves completeness by breaking down the CNN latent space into distinct and interpretable concept subspaces. However, MCD's explanations can be difficult for humans to understand, raising concerns about their practical utility. To address this, we propose Human-Understandable Multi-dimensional Concept Discovery (HU-MCD). HU-MCD uses the Segment Anything Model for concept identification and implements a CNN-specific input masking technique to reduce noise introduced by traditional masking methods. These changes to MCD, paired with the completeness relation, enable HU-MCD to enhance concept understandability while maintaining explanation faithfulness. Our experiments, including human subject studies, show that HU-MCD provides more precise and reliable explanations than existing C-XAI methods. The code is available at https://github.com/grobruegge/hu-mcd.
****

GlyphMastero: A Glyph Encoder for High-Fidelity Scene Text Editing
Tong Wang, Ting Liu, Xiaochao Qu, Chengjing Wu, Luoqi Liu, Xiaolin Hu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28523-28532
Scene text editing, a subfield of image editing, requires modifying texts in images while preserving style consistency and visual coherence with the surrounding

environment. While diffusion-based methods have shown promise in text generation, they still struggle to produce high-quality results. These methods often generate distorted or unrecognizable characters, particularly when dealing with complex characters like Chinese. In such systems, characters are composed of intricate stroke patterns and spatial relationships that must be precisely maintained. We present GlyphMastero, a specialized glyph encoder designed to guide the latent diffusion model for generating texts with stroke-level precision. Our key insight is that existing methods, despite using pretrained OCR models for feature extraction, fail to capture the hierarchical nature of text structures - from individual strokes to stroke-level interactions to overall character-level structure. To address this, our glyph encoder explicitly models and captures the cross-level interactions between local-level individual characters and global-level text lines through our novel glyph attention module. Meanwhile, our model implements a feature pyramid network to fuse the multi-scale OCR backbone features at the global-level. Through these cross-level and multi-scale fusions, we obtain more detailed glyph-aware guidance, enabling precise control over the scene text generation process. Our method achieves an 18.02% improvement in sentence accuracy over the state-of-the-art multi-lingual scene text editing baseline, while simultaneously reducing the text-region Frechet inception distance by 53.28%.
****

ConText-CIR: Learning from Concepts in Text for Composed Image Retrieval
Eric Xing, Pranavi Kolouju, Robert Pless, Abby Stylianou, Nathan Jacobs; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19638-19648
Composed image retrieval (CIR) is the task of retrieving a target image specified by a query image and a relative text that describes a semantic modification to the query image. Existing methods in CIR struggle to accurately represent the image and the text modification, resulting in subpar performance. To address this limitation, we introduce a CIR framework, \name , trained with a Text Concept-Consistency loss that encourages the representations of noun phrases in the text modification to better attend to the relevant parts of the query image. To support training with this loss function, we also propose a synthetic data generation pipeline that creates training data from existing CIR datasets or unlabeled images. We show that these components together enable stronger performance on CIR tasks, setting a new state-of-the-art in composed image retrieval in both the supervised and zero-shot settings on multiple benchmark datasets, including CIRR and CIRCO. Source code, model checkpoints, and our new datasets are available at https://github.com/mvrl/ConText-CIR.
****

MaskGaussian: Adaptive 3D Gaussian Representation from Probabilistic Masks
Yifei Liu, Zhihang Zhong, Yifan Zhan, Sheng Xu, Xiao Sun; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 681-690
While 3D Gaussian Splatting (3DGS) has demonstrated remarkable performance in novel view synthesis and real-time rendering, the high memory consumption due to the use of millions of Gaussians limits its practicality. To mitigate this issue, improvements have been made by pruning unnecessary Gaussians, either through a hand-crafted criterion or by using learned masks. However, these methods deterministically remove Gaussians based on a snapshot of the pruning moment, leading to sub-optimized reconstruction performance from a long-term perspective. To address this issue, we introduce MaskGaussian, which models Gaussians as probabilistic entities rather than permanently removing them, and utilize them according to their probability of existence. To achieve this, we propose a masked-rasterization technique that enables unused yet probabilistically existing Gaussians to receive gradients, allowing for dynamic assessment of their contribution to the evolving scene and adjustment of their probability of existence. Hence, the importance of Gaussians iteratively changes and the pruned Gaussians are selected diversely. Extensive experiments demonstrate the superiority of the proposed method in achieving better rendering quality with fewer Gaussians than previous pruning methods, pruning over 60% of Gaussians on average with only a 0.02 PSNR decline. Our code can be found at: https://github.com/kaikai23/MaskGaussian

****
Perturb-and-Revise: Flexible 3D Editing with Generative Trajectories
Susung Hong, Johanna Karras, Ricardo Martin-Brualla, Ira Kemelmacher-Shlizerman; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16293-16303
Recent advancements in text-based diffusion models have accelerated progress in 3D reconstruction and text-based 3D editing. Although existing 3D editing methods excel at modifying color, texture, and style, they struggle with extensive geometric or appearance changes, thus limiting their applications. To this end, we propose Perturb-and-Revise, which makes possible a variety of NeRF editing. First, we perturb the NeRF parameters with random initializations to create a versatile initialization. The level of perturbation is determined automatically through analysis of the local loss landscape. Then, we revise the edited NeRF via generative trajectories. Combined with the generative process, we impose identity-preserving gradients to refine the edited NeRF. Extensive experiments demonstrate that Perturb-and-Revise facilitates flexible, effective, and consistent editing of color, appearance, and geometry in 3D. For 360deg results, please visit our project page: https://susunghong.github.io/Perturb-and-Revise.
****
Birth and Death of a Rose
Chen Geng, Yunzhi Zhang, Shangzhe Wu, Jiajun Wu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26102-26113
We study the problem of generating temporal object intrinsics--temporally evolving sequences of object geometry, reflectance, and texture, such as a blooming rose--from pre-trained 2D foundation models. Unlike conventional 3D modeling and animation techniques that require extensive manual effort and expertise, we introduce a method that generates such assets with signals distilled from pretrained 2D diffusion models. To ensure the temporal consistency of object intrinsics, we propose Neural Templates for temporal-state-guided distillation, derived automatically from image features from self-supervised learning. Our method can generate high-quality temporal object intrinsics for several natural phenomena and enable the sampling and controllable rendering of these dynamic objects from any viewpoint, under any environmental lighting conditions, at any time of their lifespan. Project website: https://chen-geng.com/rose4d.
****
Learning Compatible Multi-Prize Subnetworks for Asymmetric Retrieval
Yushuai Sun, Zikun Zhou, Dongmei Jiang, Yaowei Wang, Jun Yu, Guangming Lu, Wenjie Pei; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15255-15264
Asymmetric retrieval is a typical scenario in real-world retrieval systems, where compatible models of varying capacities are deployed on platforms with different resource configurations. Existing methods generally train pre-defined networks or subnetworks with capacities specifically designed for pre-determined platforms, using compatible learning. Nevertheless, these methods suffer from limited flexibility for multi-platform deployment. For example, when introducing a new platform into the retrieval systems, developers have to train an additional model at an appropriate capacity that is compatible with existing models via backward-compatible learning. In this paper, we propose a Prunable Network with self-compatibility, which allows developers to generate compatible subnetworks at any desired capacity through post-training pruning. Thus it allows the creation of a sparse subnetwork matching the resources of the new platform without additional training. Specifically, we optimize both the architecture and weight of subnetworks at different capacities within a dense network in compatible learning. We also design a conflict-aware gradient integration scheme to handle the gradient conflicts between the dense network and subnetworks during compatible learning. Extensive experiments on diverse benchmarks and visual backbones demonstrate the effectiveness of our method. The code will be made publicly available.
****
SoundVista: Novel-View Ambient Sound Synthesis via Visual-Acoustic Binding
Mingfei Chen, Israel D. Gebru, Ishwarya Ananthabhotla, Christian Richardt, Dejan

Markovic, Jake Sandakly, Steven Krenn, Todd Keebler, Eli Shlizerman, Alexander Richard; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8331-8341

We introduce SoundVista, a method to generate the ambient sound of an arbitrary scene at novel viewpoints. Given a pre-acquired recording of the scene from sparsely distributed microphones, SoundVista can synthesize the sound of that scene from an unseen target viewpoint. The method learns the underlying acoustic transfer function that relates the signals acquired at the distributed microphones to the signal at the target viewpoint, using a limited number of known recordings. Unlike existing works, our method does not require constraints or prior knowledge of sound source details. Moreover, our method efficiently adapts to diverse room layouts, reference microphone configurations and unseen environments. To enable this, we introduce a visual-acoustic binding module that learns visual embeddings linked with local acoustic properties from panoramic RGB and depth data. We first leverage these embeddings to optimize the placement of reference microphones in any given scene. During synthesis, we leverage multiple embeddings extracted from reference locations to get adaptive weights for their contribution, conditioned on target viewpoint. We benchmark the task on both publicly available data and real-world settings. We demonstrate significant improvements over existing methods.

****

CLIP Under the Microscope: A Fine-Grained Analysis of Multi-Object Representation

Reza Abbasi, Ali Nazari, Aminreza Sefid, Mohammadali Banayeeanzade, Mohammad Hossein Rohban, Mahdieh Soleymani Baghshah; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9308-9317

Contrastive Language-Image Pre-training (CLIP) models excel in zero-shot classification, yet face challenges in complex multi-object scenarios. This study offers a comprehensive analysis of CLIP's limitations in these contexts using a specialized dataset, ComCO, designed to evaluate CLIP's encoders in diverse multi-object scenarios. Our findings reveal significant biases: the text encoder prioritizes first-mentioned objects, and the image encoder favors larger objects. Through retrieval and classification tasks, we quantify these biases across multiple CLIP variants and trace their origins to CLIP's training process, supported by analyses of the LAION dataset and training progression. Our image-text matching experiments show substantial performance drops when object size or token order changes, underscoring CLIP's instability with rephrased but semantically similar captions. Extending this to longer captions and text-to-image models like Stable Diffusion, we demonstrate how prompt order influences object prominence in generated images.

****

MetricGrids: Arbitrary Nonlinear Approximation with Elementary Metric Grids based Implicit Neural Representation

Shu Wang, Yanbo Gao, Shuai Li, Chong Lv, Xun Cai, Chuankun Li, Hui Yuan, Jinglin Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21381-21391

This paper presents MetricGrids, a novel grid-based neural representation that combines elementary metric grids in various metric spaces to approximate complex nonlinear signals. While grid-based representations are widely adopted for their efficiency and scalability, the existing feature grids with linear indexing for continuous-space points can only provide degenerate linear latent space representations, and such representations cannot be adequately compensated to represent complex nonlinear signals by the following compact decoder. To address this problem while keeping the simplicity of a regular grid structure, our approach builds upon the standard grid-based paradigm by constructing multiple elementary metric grids as high-order terms to approximate complex nonlinearities, following the Taylor expansion principle. Furthermore, we enhance model compactness with hash encoding based on different sparsities of the grids to prevent detrimental hash collisions, and a high-order extrapolation decoder to reduce explicit grid storage requirements. experimental results on both 2D and 3D reconstructions demon

strate the superior fitting and rendering accuracy of the proposed method across diverse signal types, validating its robustness and generalizability.
****

Navigating Image Restoration with VAR's Distribution Alignment Prior
Siyang Wang, Naishan Zheng, Jie Huang, Feng Zhao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7559-7569
Generative models trained on extensive high-quality datasets effectively capture the structural and statistical properties of clean images, rendering them powerful priors for transforming degraded features into clean ones in image restoration. VAR, a novel image generative paradigm, surpasses diffusion models in generation quality by applying a next-scale prediction approach. It progressively captures both global structures and fine-grained details through the autoregressive process, consistent with the multi-scale restoration principle widely acknowledged in the restoration community. Furthermore, we observe that during the image reconstruction process utilizing VAR, scale predictions automatically modulate the input, facilitating the alignment of representations at subsequent scales with the distribution of clean images. To harness VAR's adaptive distribution alignment capability in image restoration tasks, we formulate the multi-scale latent representations within VAR as the restoration prior, thus advancing our delicately designed VarFormer framework. The strategic application of these priors enables our VarFormer to achieve remarkable generalization on unseen tasks while also reducing training computational costs. Extensive experiments underscores that our VarFormer outperforms existing multi-task image restoration methods across various restoration tasks. The code is available at https://github.com/siywang541/Varformer.
****

MovieBench: A Hierarchical Movie Level Dataset for Long Video Generation
Weijia Wu, Mingyu Liu, Zeyu Zhu, Xi Xia, Haoen Feng, Wen Wang, Kevin Qinghong Lin, Chunhua Shen, Mike Zheng Shou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28984-28994
Recent advancements in video generation models, such as Stable Video Diffusion, have shown promising results, but these works primarily focus on short videos, often limited to a single scene and lacking a rich storyline. These models struggle with generating long videos that involve multiple scenes, coherent narratives, and consistent characters. Furthermore, there is currently no publicly accessible dataset specifically designed for analyzing, evaluating, and training models for long video generation. In this paper, we present MovieBench: A Hierarchical Movie-Level Dataset for Long Video Generation, which addresses these challenges by providing unique contributions: (1) character consistency across scenes, (2) long videos with rich and coherent storylines, and (3) multi-scene narratives. MovieBench features three distinct levels of annotation: the movie level, which provides a broad overview of the film; the scene level, offering a mid-level understanding of the narrative; and the shot level, which emphasizes specific moments with detailed descriptions.
****

Dissecting and Mitigating Diffusion Bias via Mechanistic Interpretability
Yingdong Shi, Changming Li, Yifan Wang, Yongxiang Zhao, Anqi Pang, Sibei Yang, Jingyi Yu, Kan Ren; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8192-8202
Diffusion models have demonstrated impressive capabilities in synthesizing diverse content. However, despite their high-quality outputs, these models often perpetuate social biases, including those related to gender and race. These biases can potentially contribute to harmful real-world consequences, reinforcing stereotypes and exacerbating inequalities in various social contexts. While existing research on diffusion bias mitigation has predominantly focused on guiding content generation, it often neglects the intrinsic mechanisms within diffusion models that causally drive biased outputs. In this paper, we investigate the internal processes of diffusion models, identifying specific decision-making mechanisms, termed bias features, embedded within the model architecture. By directly manipulating these features, our method precisely isolates and adjusts the elements r

esponsible for bias generation, permitting granular control over the bias levels in the generated content. Through experiments on both unconditional and conditional diffusion models across various social bias attributes, we demonstrate our method's efficacy in managing generation distribution while preserving image quality. We also dissect the discovered model mechanism, revealing different intrinsic features controlling fine-grained aspects of generation, boosting further research on mechanistic interpretability of diffusion models. The project website is at \href https://foundation-model-research.github.io/difflens https://foundation-model-research.github.io/difflens .
****

Graph Neural Network Combining Event Stream and Periodic Aggregation for Low-Latency Event-based Vision
Manon Dampfhoffer, Thomas Mesquida, Damien Joubert, Thomas Dalgaty, Pascal Vivet, Christoph Posch; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6909-6918
Event-based cameras asynchronously detect changes in light intensity with high temporal resolution, making them a promising alternative to RGB camera for low-latency and low-power optical flow estimation. However, state-of-the-art convolutional neural network methods create frames from the event stream, therefore losing the opportunity to exploit events for both sparse computations and low-latency prediction. On the other hand, asynchronous event graph methods could leverage both, but at the cost of avoiding any form of time accumulation, which limits the prediction accuracy. In this paper, we propose to break this accuracy-latency trade-off with a novel architecture combining an asynchronous accumulation-free event branch and a periodic aggregation branch. The periodic branch performs feature aggregations on the event graphs of past data to extract global context information, which improves accuracy without introducing any latency.The solution could predict optical flow per event with a latency of tens of microseconds on asynchronous hardware, which represents a gain of three orders of magnitude with respect to state-of-the-art frame-based methods, with 48x less operations per second. We show that the solution can detect rapid motion changes faster than a periodic output. This work proposes, for the first time, an effective solution for ultra low-latency and low-power optical flow prediction from event cameras.
****

Be More Specific: Evaluating Object-centric Realism in Synthetic Images
Anqi Liang, Ciprian Corneanu, Qianli Feng, Giorgio Giannone, Aleix Martinez; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28842-28851
Evaluation of synthetic images is important for both model development and selection. An ideal evaluation should be specific, accurate and aligned with human perception. This paper addresses the problem of evaluating realism of objects in synthetic images. Although methods has been proposed to evaluate holistic realism, there are no methods tailored towards object-centric realism evaluation. In this work, we define a new standard for assessing object-centric realism that follows a shape-texture breakdown and proposes the first object-centric realism evaluation dataset for synthetic images. The dataset contains images generated from state-of-the-art image generative models and is richly annotated at object level across a diverse set of object categories. We then design and train the OLIP model, a dedicated architecture that considerably outperforms any existing baseline on object-centric realism evaluation.
****

Correlative and Discriminative Label Grouping for Multi-Label Visual Prompt Tuning
Lei-Lei Ma, Shuo Xu, Ming-Kun Xie, Lei Wang, Dengdi Sun, Haifeng Zhao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25434-25443
Modeling label correlations has always played a pivotal role in multi-label image classification (MLC), attracting significant attention from researchers. However, recent studies have overemphasized co-occurrence relationships among labels, which can lead to overfitting risk on this overemphasis, resulting in suboptima

l models. To tackle this problem, we advocate for balancing correlative and disc
riminative relationships among labels to mitigate the risk of overfitting and en
hance model performance. To this end, we propose the Multi-Label Visual Prompt T
uning framework, a novel and parameter-efficient method that groups classes into
 multiple class subsets according to label co-occurrence and mutual exclusivity
relationships, and then models them respectively to balance the two relationship
s. In this work, since each group contains multiple classes, multiple prompt tok
ens are adopted within Vision Transformer (ViT) to capture the correlation or di
scriminative label relationship within each group, and effectively learn correla
tion or discriminative representations for class subsets. On the other hand, eac
h group contains multiple group-aware visual representations that may correspond
 to multiple classes, and the mixture of experts (MoE) model can cleverly assign
 them from the group-aware to the label-aware, adaptively obtaining label-aware
representation, which is more conducive to classification. Experiments on multip
le benchmark datasets show that our proposed approach achieves competitive resul
ts and outperforms SOTA methods on multiple pre-trained models.
****

LoRACLR: Contrastive Adaptation for Customization of Diffusion Models
Enis Simsar, Thomas Hofmann, Federico Tombari, Pinar Yanardag; Proceedings of th
e Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13189-131
98
Recent advances in text-to-image customization have enabled high-fidelity, conte
xt-rich generation of personalized images, allowing specific concepts to appear
in a variety of scenarios. However, current methods struggle with combining mult
iple personalized models, often leading to attribute entanglement or requiring s
eparate training to preserve concept distinctiveness. We present LoRACLR, a nove
l approach for multi-concept image generation that merges multiple LoRA models,
each fine-tuned for a distinct concept, into a single, unified model without add
itional individual fine-tuning. LoRACLR uses a contrastive objective to align an
d merge the weight spaces of these models, ensuring compatibility while minimizi
ng interference. By enforcing distinct yet cohesive representations for each con
cept, LoRACLR enables efficient, scalable model composition for high-quality, mu
lti-concept image synthesis. Our results highlight the effectiveness of LoRACLR
in accurately merging multiple concepts, advancing the capabilities of personali
zed image generation.
****

ArtFormer: Controllable Generation of Diverse 3D Articulated Objects
Jiayi Su, Youhe Feng, Zheng Li, Jinhua Song, Yangfan He, Botao Ren, Botian Xu; P
roceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 202
5, pp. 1894-1904
This paper presents a novel framework for modeling and conditional generation of
 3D articulated objects. Troubled by flexibility-quality tradeoffs, existing met
hods are often limited to using predefined structures or retrieving shapes from
static datasets. To address these challenges, we parameterize an articulated obj
ect as a tree of tokens and employ a transformer to generate both the object's h
igh-level geometry code and its kinematic relations. Subsequently, each sub-part
's geometry is further decoded using a signed-distance-function (SDF) shape prio
r, facilitating the synthesis of high-quality 3D shapes. Our approach enables th
e generation of diverse objects with high-quality geometry and varying number of
 parts. Comprehensive experiments on conditional generation from text descriptio
ns demonstrate the effectiveness and flexibility of our method.
****

Opportunistic Single-Photon Time of Flight
Sotiris Nousias, Mian Wei, Howard Xiao, Maxx Wu, Shahmeer Athar, Kevin J. Wang,
Anagh Malik, David A. Barmherzig, David B. Lindell, Kyros N. Kutulakos; Proceedi
ngs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp.
15852-15862
Scattered light from pulsed lasers is increasingly part of our ambient illuminat
ion, as many devices rely on them for active 3D sensing. In this work, we ask: c
an these "ambient" light signals be detected and leveraged for passive 3D vision

? We show that pulsed lasers, despite being weak and fluctuating at MHz to GHz f
requencies, leave a distinctive sinc comb pattern in the temporal frequency doma
in of incident flux that is specific to each laser and invariant to the scene. T
his enables their passive detection and analysis with a free-running SPAD camera
, even when they are unknown, asynchronous, out of sight, and emitting concurren
tly. We show how to synchronize with such lasers computationally, characterize t
heir pulse emissions, separate their contributions, and--if many are present--lo
calize them in 3D and recover a depth map of the camera's field of view. We use
our camera prototype to demonstrate (1) a first-of-its-kind visualization of asy
nchronously propagating light pulses from multiple lasers through the same scene
, (2) passive estimation of a laser's MHz-scale pulse repetition frequency with
mHz precision, and (3) mm-scale 3D imaging over room-scale distances by passivel
y harvesting photons from two or more out-of-view lasers.
****

Bridging Gait Recognition and Large Language Models Sequence Modeling
Shaopeng Yang, Jilong Wang, Saihui Hou, Xu Liu, Chunshui Cao, Liang Wang, Yongzh
en Huang; Proceedings of the Computer Vision and Pattern Recognition Conference
(CVPR), 2025, pp. 3460-3469
Gait sequences exhibit sequential structures and contextual relationships simila
r to those in natural language, where each element--whether a word or a gait ste
p--is connected to its predecessors and successors. This similarity enables the
transformation of gait sequences into "texts" containing identity-related inform
ation. Large Language Models (LLMs), designed to understand and generate sequent
ial data, can thus be utilized for gait sequence modeling to enhance gait recogn
ition performance. Leveraging these insights, we make a pioneering effort to app
ly LLMs to gait recognition, which we refer to as GaitLLM. Specifically, we prop
ose the Gait-to-Language (G2L) module, which converts gait sequences into a text
ual format suitable for LLMs, and the Language-to-Gait (L2G) module, which maps
the LLM's output back to the gait feature space, thereby bridging the gap betwee
n LLM outputs and gait recognition. Notably, GaitLLM leverages the powerful mode
ling capabilities of LLMs without relying on complex architectural designs, impr
oving gait recognition performance with only a small number of trainable paramet
ers. Our method achieves state-of-the-art results on four popular gait datasets-
-SUSTech1K, CCPG, Gait3D, and GREW--demonstrating the effectiveness of applying
LLMs in this domain. This work highlights the potential of LLMs to significantly
enhance gait recognition, paving the way for future research and practical appl
ications.
****

Argus: Vision-Centric Reasoning with Grounded Chain-of-Thought
Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Ja
n Kautz, Yu-Xiong Wang, Zhiding Yu; Proceedings of the Computer Vision and Patte
rn Recognition Conference (CVPR), 2025, pp. 14268-14280
Recent advances in multimodal large language models (MLLMs) have demonstrated re
markable capabilities in vision-language tasks, yet they often struggle with vis
ion-centric scenarios where precise visual focus is needed for accurate reasonin
g. In this paper, we introduce Argus to address these limitations with a new vis
ual attention grounding mechanism. Our approach employs object-centric grounding
as visual chain-of-thought signals, enabling more effective goal-conditioned vi
sual attention during multimodal reasoning tasks. Evaluations on diverse benchma
rks demonstrate that Argus excels in both multimodal reasoning tasks and referri
ng object grounding tasks. Extensive analysis further validates various design c
hoices of Argus, and reveals the effectiveness of explicit language-guided visua
l region-of-interest engagement in MLLMs, highlighting the importance of advanci
ng multimodal intelligence from a visual-centric perspective.
****

Bootstrap Your Own Views: Masked Ego-Exo Modeling for Fine-grained View-invarian
t Video Representations
Jungin Park, Jiyoung Lee, Kwanghoon Sohn; Proceedings of the Computer Vision and
Pattern Recognition Conference (CVPR), 2025, pp. 13661-13670
View-invariant representation learning from egocentric (first-person, ego) and e

xocentric (third-person, exo) videos is a promising approach toward generalizing video understanding systems across multiple viewpoints. However, this area has been underexplored due to the substantial differences in perspective, motion patterns, and context between ego and exo views. In this paper, we propose a novel masked ego-exo modeling that promotes both causal temporal dynamics and cross-view alignment, called Bootstrap Your Own Views (BYOV), for fine-grained view-invariant video representation learning from unpaired ego-exo videos. We highlight the importance of capturing the compositional nature of human actions as a basis for robust cross-view understanding. Specifically, self-view masking and cross-view masking predictions are designed to learn view-invariant and powerful representations concurrently. Experimental results demonstrate that our BYOV significantly surpasses existing approaches with notable gains across all metrics in four downstream ego-exo video tasks. The code is available at https://github.com/park-jungin/byov

****

SFDM: Robust Decomposition of Geometry and Reflectance for Realistic Face Rendering from Sparse-view Images

Daisheng Jin, Jiangbei Hu, Baixin Xu, Yuxin Dai, Chen Qian, Ying He; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26409-26419

In this study, we introduce a novel two-stage technique for decomposing and reconstructing facial features from sparse-view images, a task made challenging by the unique geometry and complex skin reflectance of each individual. To synthesize 3D facial models more realistically, we endeavor to decouple key facial attributes from the RGB color, including geometry, diffuse reflectance, and specular reflectance. Specifically, we design a Sparse-view Face Decomposition Model (SFDM): 1) In the first stage, we create a general facial template from a wide array of individual faces, encapsulating essential geometric and reflectance characteristics. 2) Guided by this template, we refine a specific facial model for each individual in the second stage, considering the interaction between geometry and reflectance, as well as the effects of subsurface scattering on the skin. With these advances, our method can reconstruct high-quality facial representations from as few as three images. The comprehensive evaluation and comparison reveal that our approach outperforms existing methods by effectively disentangling geometric and reflectance components, significantly enhancing the quality of synthesized novel views, and paving the way for applications in facial relighting and reflectance editing.

****

DiSRT-In-Bed: Diffusion-Based Sim-to-Real Transfer Framework for In-Bed Human Mesh Recovery

Jing Gao, Ce Zheng, Laszlo A. Jeni, Zackory Erickson; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1829-1838

In-bed human mesh recovery can be crucial and enabling for several healthcare applications, including sleep pattern monitoring, rehabilitation support, and pressure ulcer prevention. However, it is difficult to collect large real-world visual datasets in this domain, in part due to privacy and expense constraints, which in turn presents significant challenges for training and deploying deep learning models. Existing in-bed human mesh estimation methods often rely heavily on real-world data, limiting their ability to generalize across different in-bed scenarios, such as varying coverings and environmental settings. To address this, we propose a Sim-to-Real Transfer Framework for in-bed human mesh recovery from overhead depth images, which leverages large-scale synthetic data alongside limited or no real-world samples. We introduce a diffusion model that bridges the gap between synthetic data and real data to support generalization in real-world in-bed pose and body inference scenarios. Extensive experiments and ablation studies validate the effectiveness of our framework, demonstrating significant improvements in robustness and adaptability across diverse healthcare scenarios. Project page can be found at https://jing-g2.github.io/DiSRT-In-Bed/.

****

Ouroboros3D: Image-to-3D Generation via 3D-aware Recursive Diffusion

Hao Wen, Zehuan Huang, Yaohui Wang, Xinyuan Chen, Lu Sheng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21631-21641
Existing image-to-3D creation methods typically split the task into two individual stage: multi-view image generation and 3D reconstruction, leading to two main limitations: (1) In multi-view generation stage, the multi-view generated images present a challenge to preserving 3D consistency;; (2) In 3D reconstruction stage, there is a domain gap between real training data and generated multi-view input during inference. To address these issues, we propose Ouroboros3D, end-to-end trainable framework that integrates multi-view generation and 3D reconstruction into a recursive diffusion process through feedback mechanism.Our framework operates through iterative cycles where each cycle consists of a denoising process and a reconstruction step.By incorporating a 3D-aware feedback mechanism, our multi-view generative model leverages the explicit 3D geometric information (e.g. texture, position) from the feedback of reconstruction results of the previous process as conditions, thus modeling consistency at the 3D geometric level. Furthermore, through joint training of both the multi-view generative and reconstruction models, we alleviate reconstruction stage domain gap and enable mutual enhancement within the recursive process. Experimental results demonstrate that Ouroboros3D outperforms methods that treat these stages separately and those that combine them only during inference, achieving superior multi-view consistency and producing 3D models with higher geometric realism.
****

QMambaBSR: Burst Image Super-Resolution with Query State Space Model
Xin Di, Long Peng, Peizhe Xia, Wenbo Li, Renjing Pei, Yang Cao, Yang Wang, Zheng-Jun Zha; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23080-23090
Burst super-resolution (BurstSR) aims to reconstruct high-resolution images by fusing subpixel details from multiple low-resolution burst frames. The primary challenge lies in effectively extracting useful information while mitigating the impact of high-frequency noise. Most existing methods rely on frame-by-frame fusion, which often struggles to distinguish informative subpixels from noise, leading to suboptimal performance. To address these limitations, we introduce a novel Query Mamba Burst Super-Resolution (QMambaBSR) network. Specifically, we observe that sub-pixels have consistent spatial distribution while noise appears randomly. Considering the entire burst sequence during fusion allows for more reliable extraction of consistent subpixels and better suppression of noise outliers. Based on this, a Query State Space Model (QSSM) is proposed for both inter-frame querying and intra-frame scanning, enabling a more efficient fusion of useful subpixels. Additionally, to overcome the limitations of static upsampling methods that often result in over-smoothing, we propose an Adaptive Upsampling (AdaUp) module that dynamically adjusts the upsampling kernel to suit the characteristics of different burst scenes, achieving superior detail reconstruction. Extensive experiments on four benchmark datasets--spanning both synthetic and real-world images--demonstrate that QMambaBSR outperforms existing state-of-the-art methods.
****

Encapsulated Composition of Text-to-Image and Text-to-Video Models for High-Quality Video Synthesis
Tongtong Su, Chengyu Wang, Bingyan Liu, Jun Huang, Dongming Lu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18209-18218
In recent years, large text-to-video (T2V) synthesis models have garnered considerable attention for their abilities to generate videos from textual descriptions. However, achieving both high imaging quality and effective motion representation remains a significant challenge for these T2V models. Existing approaches often adapt pre-trained text-to-image (T2I) models to refine video frames, leading to issues such as flickering and artifacts due to inconsistencies across frames. In this paper, we introduce EVS, a training-free \underline E ncapsulated \underline V ideo \underline S ynthesizer that composes T2I and T2V models to enhance both visual fidelity and motion smoothness of generated videos. Our approach utilizes a well-trained diffusion-based T2I model to refine low-quality video fra

mes by treating them as out-of-distribution samples, effectively optimizing them with noising and denoising steps. Meanwhile, we employ T2V backbones to ensure consistent motion dynamics. By encapsulating the T2V temporal-only prior into the T2I generation process, EVS successfully leverages the strengths of both types of models, resulting in videos of improved imaging and motion quality. Experimental results validate the effectiveness of our approach compared to previous approaches.Our composition process also leads to a significant improvement of 1.6x-4.5x speedup in inference time.

****

Multi-Group Proportional Representations for Text-to-Image Models

Sangwon Jung, Alex Oesterling, Claudio Mayrink Verdun, Sajani Vithana, Taesup Moon, Flavio P. Calmon; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23744-23754

Text-to-image (T2I) generative models can create vivid, realistic images from textual descriptions. As these models proliferate, they expose new concerns about their ability to represent diverse demographic groups, propagate stereotypes, and efface minority populations. Despite growing attention to the "safe" and "responsible" design of artificial intelligence (AI), there is no established methodology to systematically measure and control representational harms in image generation. This paper introduces a novel framework to measure the representation of intersectional groups in images generated by T2I models by applying the Multi-Group Proportional Representation (MPR) metric. MPR evaluates the worst-case deviation of representation statistics across given population groups in images produced by a generative model, allowing for flexible and context-specific measurements based on user requirements. We also develop an algorithm to optimize T2I models for this metric. Through experiments, we demonstrate that MPR can effectively measure representation statistics across multiple intersectional groups and, when used as a training objective, can guide models toward a more balanced generation across demographic groups while maintaining generation quality.

****

Towards Generalizable Trajectory Prediction using Dual-Level Representation Learning and Adaptive Prompting

Kaouther Messaoud, Matthieu Cord, Alexandre Alahi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27564-27574

Existing vehicle trajectory prediction models struggle with generalizability, prediction uncertainties, and handling complex interactions. It is often due to limitations like complex architectures customized for a specific dataset and inefficient multimodal handling. We propose Perceiver with Register queries (PerReg+), a novel trajectory prediction framework that introduces: (1) Dual-Level Representation Learning via Self-Distillation (SD) and Masked Reconstruction (MR), capturing global context and fine-grained details. Additionally, our approach of reconstructing segment-level trajectories and lane segments from masked inputs with query drop, enables effective use of contextual information and improves generalization; (2) Enhanced Multimodality using register-based queries and pretraining, eliminating the need for clustering and suppression; and (3) Adaptive Prompt Tuning during fine-tuning, freezing the main architecture and optimizing a small number of prompts for efficient adaptation. PerReg+ sets a new state-of-the-art performance on nuScenes, Argoverse 2, and Waymo Open Motion Dataset (WOMD). Remarkable, our pretrained model reduces the error by 6.8% on smaller datasets, and multi-dataset training enhances generalization. In cross-domain tests, PerReg+ reduces B-FDE by 11.8% compared to its non-pretrained variant.

****

CoMatcher: Multi-View Collaborative Feature Matching

Jintao Zhang, Zimin Xia, Mingyue Dong, Shuhan Shen, Linwei Yue, Xianwei Zheng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21970-21980

This paper proposes a multi-view collaborative matching strategy for reliable track construction in complex scenarios. We observe that the pairwise matching paradigms applied to image set matching often result in ambiguous estimation when the selected independent pairs exhibit significant occlusions or extreme viewpoin

t changes. This challenge primarily stems from the inherent uncertainty in inter preting intricate 3D structures based on limited two-view observations, as the 3 D-to-2D projection leads to significant information loss. To address this, we in troduce CoMatcher, a deep multi-view matcher to (i) leverage complementary conte xt cues from different views to form a holistic 3D scene understanding and (ii) utilize cross-view projection consistency to infer a reliable global solution. B uilding on CoMatcher, we develop a groupwise framework that fully exploits cross -view relationships for large-scale matching tasks. Extensive experiments on var ious complex scenarios demonstrate the superiority of our method over the mainst ream two-view matching paradigm.

****

COUNTS: Benchmarking Object Detectors and Multimodal Large Language Models under Distribution Shifts

Jiansheng Li, Xingxuan Zhang, Hao Zou, Yige Guo, Renzhe Xu, Yilong Liu, Chuzhao Zhu, Yue He, Peng Cui; Proceedings of the Computer Vision and Pattern Recognitio n Conference (CVPR), 2025, pp. 9186-9198

Current object detectors often suffer significant performance degradation in rea l-world applications when encountering distributional shifts, posing serious ris ks in high-stakes domains such as autonomous driving and medical diagnosis. Cons equently, the out-of-distribution (OOD) generalization capability of object dete ctors has garnered increasing attention from researchers. Despite this growing i nterest, there remains a lack of a large-scale, comprehensive dataset and evalua tion benchmark with fine-grained annotations tailored to assess the OOD generali zation on more intricate tasks like object detection and grounding. To address t his gap, we introduce COUNTS, a large-scale OOD dataset with object-level annota tions. COUNTS encompasses 14 natural distributional shifts, over 222K samples, a nd more than 1,196K labeled bounding boxes. Leveraging COUNTS, we introduce two novel benchmarks: O(OD) and OODG. OODOD is designed to comprehensively evaluate the OOD generalization capabilities of object detectors by utilizing controlled distribution shifts between training and testing data. OODG, on the other hand, aims to assess the OOD generalization of grounding abilities in multimodal large language models (MLLMs). Our findings reveal that, while large models and exten sive pre-training data substantially enhance performance in in-distribution (IID) scenarios, significant limitations and opportunities for improvement persist i n OOD contexts for both object detectors and MLLMs. In visual grounding tasks, e ven the advanced GPT-4o and Gemini-1.5 only achieve 56.7% and 28.0% accuracy, re spectively. We hope COUNTS facilitates advancements in the development and asses sment of robust object detectors and MLLMs capable of maintaining high performan ce under distributional shifts.

****

Retrieving Semantics from the Deep: an RAG Solution for Gesture Synthesis

M. Hamza Mughal, Rishabh Dabral, Merel C.J. Scholman, Vera Demberg, Christian Th eobalt; Proceedings of the Computer Vision and Pattern Recognition Conference (C VPR), 2025, pp. 16578-16588

Non-verbal communication often comprises of semantically rich gestures that help convey the meaning of an utterance. Producing such semantic co-speech gestures has been a major challenge for the existing neural systems that can generate rhy thmic beat gestures, but struggle to produce semantically meaningful gestures. T herefore, we present RAG-Gesture, a diffusion-based gesture generation approach that leverages Retrieval Augmented Generation (RAG) to produce natural-looking a nd semantically rich gestures. Our neuro-explicit gesture generation approach is designed to produce semantic gestures grounded in interpretable linguistic know ledge. We achieve this by using explicit domain knowledge to retrieve exemplar m otions from a database of co-speech gestures. Once retrieved, we then inject the se semantic exemplar gestures into our diffusion-based gesture generation pipeli ne using DDIM inversion and retrieval guidance at the inference time without any need of training. Further, we propose a control paradigm for guidance, that all ows the users to modulate the amount of influence each retrieval insertion has o ver the generated sequence. Our comparative evaluations demonstrate the validity of our approach against recent gesture generation approaches. The reader is urg

ed to explore the results on https://vcai.mpi-inf.mpg.de/projects/RAG-Gesture/
****
HOT: Hadamard-based Optimized Training
Seonggon Kim, Juncheol Shin, Seung-taek Woo, Eunhyeok Park; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4787-4796
It has become increasingly important to optimize backpropagation to reduce memory usage and computational overhead. Achieving this goal is highly challenging, as multiple objectives must be considered jointly while maintaining training quality. In this paper, we focus on matrix multiplication, which accounts for the largest portion of training costs, and analyze its backpropagation in detail to identify lightweight techniques that offer the best benefits. Based on this analysis, we introduce a novel method, Hadamard-based Optimized Training (HOT). In this approach, we apply Hadamard-based optimizations, such as Hadamard quantization and Hadamard low-rank approximation, selectively and with awareness of the suitability of each optimization for different backward paths. Additionally, we introduce two enhancements: activation buffer compression and layer-wise quantizer selection. Our extensive analysis shows that HOT achieves up to 75% memory savings and a 2.6 times acceleration on real GPUs, with negligible accuracy loss compared to FP32 precision.
****
Towards a Universal Synthetic Video Detector: From Face or Background Manipulations to Fully AI-Generated Content
Rohit Kundu, Hao Xiong, Vishal Mohanty, Athula Balachandran, Amit K. Roy-Chowdhury; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28050-28060
Existing DeepFake detection techniques primarily focus on facial manipulations, such as face-swapping or lip-syncing. However, advancements in text-to-video (T2V) and image-to-video (I2V) generative models now allow fully AI-generated synthetic content and seamless background alterations, challenging face-centric detection methods and demanding more versatile approaches.To address this, we introduce the Universal Network for Identifying Tampered and Engineered videos (UNITE) model, which, unlike traditional detectors, captures full-frame manipulations. UNITE extends detection capabilities to scenarios without faces, non-human subjects, and complex background modifications. It leverages a transformer-based architecture that processes domain-agnostic features extracted from videos via the SigLIP-So400M foundation model. Given limited datasets encompassing both facial/background alterations and T2V/I2V content, we integrate task-irrelevant data alongside standard DeepFake datasets in training. We further mitigate the model's tendency to over-focus on faces by incorporating an attention-diversity (AD) loss, which promotes diverse spatial attention across video frames. Combining AD loss with cross-entropy improves detection performance across varied contexts. Comparative evaluations demonstrate that UNITE outperforms state-of-the-art detectors on datasets (in cross-data settings) featuring face/background manipulations and fully synthetic T2V/I2V videos, showcasing its adaptability and generalizable detection capabilities.
****
TokenFlow: Unified Image Tokenizer for Multimodal Understanding and Generation
Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K. Du, Zehuan Yuan, Xinglong Wu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2545-2555
We present TokenFlow, a novel unified image tokenizer that bridges the long-standing gap between multimodal understanding and generation. Prior research attempt to employ a single reconstruction-targeted Vector Quantization (VQ) encoder for unifying these two tasks. We observe that understanding and generation require fundamentally different granularities of visual information. This leads to a critical trade-off, particularly compromising performance in multimodal understanding tasks. TokenFlow addresses this challenge through an innovative dual-codebook architecture that decouples semantic and pixel-level feature learning while maintaining their alignment via a shared mapping mechanism. This design enables direct access to both high-level semantic representations crucial for understandin

g tasks and fine-grained visual features essential for generation through shared indices. Our extensive experiments demonstrate TokenFlow's superiority across multiple dimensions. Leveraging TokenFlow, we demonstrate for the first time that discrete visual input can surpass LLaVA-1.5 13B in understanding performance, achieving a 7.2% average improvement. For image reconstruction, we achieve a strong FID score of 0.63 at 384x384 resolution. Moreover, TokenFlow establishes state-of-the-art performance in autoregressive image generation with a GenEval score of 0.55 at 256x256 resolution, achieving comparable results to SDXL. Our code and models are released at https://github.com/ByteFlow-AI/TokenFlow.
****

Improving Personalized Search with Regularized Low-Rank Parameter Updates
Fiona Ryan, Josef Sivic, Fabian Caba Heilbron, Judy Hoffman, James M. Rehg, Bryan Russell; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19748-19757
Personalized vision-language retrieval seeks to recognize new concepts (e.g. "my dog Fido") from only a few examples. This task is challenging because it requires not only learning a new concept from a few images, but also integrating the personal and general knowledge together to recognize the concept in different contexts. In this paper, we show how to effectively adapt the internal representation of a vision-language dual encoder model for personalized vision-language retrieval. We find that regularized low-rank adaption of a small set of parameters in the language encoder's final layer serves as a highly effective alternative to textual inversion for recognizing the personal concept while preserving general knowledge. Additionally, we explore strategies for combining parameters of multiple learned personal concepts, finding that parameter addition is effective. To evaluate how well general knowledge is preserved in a finetuned representation, we introduce a metric that measures image retrieval accuracy based on captions generated by a vision language model (VLM). Our approach achieves state-of-the-art accuracy on two benchmarks for personalized image retrieval with natural language queries - DeepFashion2 and ConConChi - outperforming the prior art by 4%-22% on personal retrievals.
****

A Focused Human Body Model for Accurate Anthropometric Measurements Extraction
Shuhang Chen, Xianliang Huang, Zhizhou Zhong, Juhong Guan, Shuigeng Zhou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22658-22667
3D anthropometric measurements have a variety of applications in industrial design and architecture (e.g. vehicle seating and cockpits), Clothing (e.g. military uniforms), Ergonomics (e.g. seating) and Medicine (e.g. nutrition and diabetes) etc. Therefore, there is a need for systems that can accurately extract human body measurements. Current methods estimate human body measurements from 3D scans, resulting in a heavy data collection burden. Moreover, minor variations in camera angle, distance, and body postures may significantly affect the measurement accuracy. In response to these challenges, this paper introduces a focused human body model for accurately extracting anthropometric measurements. Concretely, we design a Bypass Network based on CNN and ResNet architectures, which augments the frozen backbone SMPLer-X with additional feature extraction capabilities. On the other hand, to boost the efficiency of training a large-scale model, we integrate a dynamical loss function that automatically recalibrates the weights to make the network focus on targeted anthropometric parts. In addition, we construct a multimodal body measurement benchmark dataset consisting of depth, point clouds, mesh and corresponding body measurements to support model evaluation and future anthropometric measurement research. Extensive experiments on both open-source and the proposed human body datasets demonstrate the superiority of our approach over existing counterparts, including the current mainstream commercial body measurement software.
****

SnapGen-V: Generating a Five-Second Video within Five Seconds on a Mobile Device
Yushu Wu, Zhixing Zhang, Yanyu Li, Yanwu Xu, Anil Kag, Yang Sui, Huseyin Coskun, Ke Ma, Aleksei Lebedev, Ju Hu, Dimitris N. Metaxas, Yanzhi Wang, Sergey Tulyako

v, Jian Ren; Proceedings of the Computer Vision and Pattern Recognition Conferen ce (CVPR), 2025, pp. 2479-2490
We have witnessed the unprecedented success of diffusion-based video generation over the past year. Recently proposed models from the community have wielded the power to generate cinematic and high-resolution videos with smooth motions from arbitrary input prompts. However, as a supertask of image generation, video gen eration models require more computation and are thus hosted mostly on cloud serv ers, limiting broader adoption among content creators. In this work, we propose a comprehensive acceleration framework to bring the power of the large-scale vid eo diffusion model to the hands of edge users. From the network architecture sco pe, we initialize from a compact image backbone and search out the design and ar rangement of temporal layers to maximize hardware efficiency. In addition, we pr opose a dedicated adversarial fine-tuning algorithm for our efficient model and reduce the denoising steps to 4. Our model, with only 0.6B parameters, can gener ate a 5-second video on an iPhone 16 PM within 5 seconds. Compared to server-sid e models that take minutes on powerful GPUs to generate a single video, we accel erate the generation by magnitudes while delivering on-par quality.
****

Adapting Dense Matching for Homography Estimation with Grid-based Acceleration
Kaining Zhang, Yuxin Deng, Jiayi Ma, Paolo Favaro; Proceedings of the Computer V ision and Pattern Recognition Conference (CVPR), 2025, pp. 6294-6303
Current deep homography estimation methods are typically constrained to processi ng low-resolution image pairs due to network architecture and computational limi tations. For high-resolution images, downsampling is often required, which can g reatly degrade estimation accuracy. In contrast, image matching methods, which m atch pixels and compute homography from correspondences, provide greater resolut ion flexibility. So in this work, we revisit the traditional image matching para digm for homography estimation and propose GFNet, a Grid Flow regression Networ k that adapts the high-accuracy dense matching framework for homography estimati on while enhancing efficiency through a grid-based strategy--estimating flow onl y over a coarse grid by leveraging homography's global smoothness. We demonstrat e the effectiveness of GFNet on a wide range of experiments on multiple datasets , including the common scene MSCOCO, multimodal datasets VIS-IR and GoogleMap, a nd the dynamic scene VIRAT. Notably, on 448x448 GoogleMap, GFNet achieves an imp rovement of +13.5% in auc@3 while reducing MACs by ~47% compared to the SOTA den se matching method. Additionally, it shows a 1.8ximprovement in auc@3 over the S OTA deep homography method. Code is available at \textcolor[rgb] 0.95, 0.08, 0.5 8 https://github.com/KN-Zhang/GFNet .
****

HyperLoRA: Parameter-Efficient Adaptive Generation for Portrait Synthesis
Mengtian Li, Jinshu Chen, Wanquan Feng, Bingchuan Li, Fei Dai, Songtao Zhao, Qia n He; Proceedings of the Computer Vision and Pattern Recognition Conference (CVP R), 2025, pp. 13114-13123
Personalized portrait synthesis, essential in domains like social entertainment, has recently made significant progress. Person-wise fine-tuning based methods, such as LoRA and DreamBooth, can produce photorealistic outputs but need trainin g on individual samples, consuming time and resources and posing an unstable ris k. Adapter based techniques such as IP-Adapter freeze the foundational model par ameters and employ a plug-in architecture to enable zero-shot inference, but the y often exhibit a lack of naturalness and authenticity, which are not to be over looked in portrait synthesis tasks. In this paper, we introduce a parameter-effi cient adaptive generation method, namely HyperLoRA, that uses an adaptive plug-i n network to generate LoRA weights, merging the superior performance of LoRA wit h the zero-shot capability of adapter scheme. Through our carefully designed net work structure and training strategy, we achieve zero-shot personalized portrait generation (supporting both single and multiple image inputs) with high photore alism, fidelity, and editability.
****

ACE: Anti-Editing Concept Erasure in Text-to-Image Models
Zihao Wang, Yuxiang Wei, Fan Li, Renjing Pei, Hang Xu, Wangmeng Zuo; Proceedings

of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23505-23515
Recent advance in text-to-image diffusion models have significantly facilitated the generation of high-quality images, but also raising concerns about the illegal creation of harmful content, such as copyrighted images. Existing concept erasure methods achieve superior results in preventing the production of erased concept from prompts, but typically perform poorly in preventing undesired editing. To address this issue, we propose an Anti-Editing Concept Erasure (ACE) method, which not only erases the target concept during generation but also filters out it during editing. Specifically, we propose to inject the erasure guidance into both conditional and the unconditional noise prediction, enabling the model to effectively prevent the creation of erasure concepts during both editing and generation. Furthermore, a stochastic correction guidance is introduced during training to address the erosion of unrelated concepts. We conducted erasure editing experiments with representative editing methods (i.e., LEDITS++ and MasaCtrl) to erase IP characters, and the results indicate that our ACE effectively filters out target concepts in both types of edits. Additional experiments on erasing explicit concepts and artistic styles further demonstrate that our ACE performs favorably against state-of-the-art methods. Our code will be publicly available at https://github.com/120L020904/ACE.
****

EchoMatch: Partial-to-Partial Shape Matching via Correspondence Reflection
Yizheng Xie, Viktoria Ehm, Paul Roetzer, Nafie El Amrani, Maolin Gao, Florian Bernard, Daniel Cremers; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11665-11675
Finding correspondences between 3D shapes is a crucial problem in computer vision and graphics. While most research has focused on finding correspondences in settings where at least one of the shapes is complete, the realm of partial-to-partial shape matching remains under-explored. Yet, it is important since in many applications shapes are only observed partially due to occlusion or scanning. Finding correspondences between partial shapes comes with an additional challenge: We not only want to identify correspondences between points on either shape but also have to determine which points of each shape actually have a partner. To tackle this challenging problem, we present EchoMatch, a novel framework for partial-to-partial shape matching that incorporates the concept of correspondence reflection to enable an overlap prediction within a functional map framework. With this approach, we show that we can outperform current SOTA methods in challenging partial-to-partial shape matching problems. Our code is available at https://echo-match.github.io.
****

CoSDH: Communication-Efficient Collaborative Perception via Supply-Demand Awareness and Intermediate-Late Hybridization
Junhao Xu, Yanan Zhang, Zhi Cai, Di Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6834-6843
Multi-agent collaborative perception enhances perceptual capabilities by utilizing information from multiple agents and is considered a fundamental solution to the problem of weak single-vehicle perception in autonomous driving. However, existing collaborative perception methods face a dilemma between communication efficiency and perception accuracy. To address this issue, we propose a novel communication-efficient collaborative perception framework based on supply-demand awareness and intermediate-late hybridization, dubbed as CoSDH. By modeling the supply-demand relationship between agents, the framework refines the selection of collaboration regions, reducing unnecessary communication cost while maintaining accuracy. In addition, we innovatively introduce the intermediate-late hybrid collaboration mode, where late-stage collaboration compensates for the performance degradation in collaborative perception under low communication bandwidth. Extensive experiments on multiple datasets, including both simulated and real-world scenarios, demonstrate that CoSDH achieves state-of-the-art detection accuracy and optimal bandwidth trade-offs, delivering superior detection precision under real communication bandwidths, thus proving its effectiveness and practical appli

cability. The code will be released at https://github.com/Xu2729/CoSDH.
****

Stereo Anywhere: Robust Zero-Shot Deep Stereo Matching Even Where Either Stereo or Mono Fail

Luca Bartolomei, Fabio Tosi, Matteo Poggi, Stefano Mattoccia; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1013-1027

We introduce Stereo Anywhere, a novel stereo-matching framework that combines geometric constraints with robust priors from monocular depth Vision Foundation Models (VFMs). By elegantly coupling these complementary worlds through a dual-branch architecture, we seamlessly integrate stereo matching with learned contextual cues. Following this design, our framework introduces novel cost volume fusion mechanisms that effectively handle critical challenges such as textureless regions, occlusions, and non-Lambertian surfaces. Through our novel optical illusion dataset, MonoTrap, and extensive evaluation across multiple benchmarks, we demonstrate that our synthetic-only trained model achieves state-of-the-art results in zero-shot generalization, significantly outperforming existing solutions while showing remarkable robustness to challenging cases such as mirrors and transparencies.
****

Order-Robust Class Incremental Learning: Graph-Driven Dynamic Similarity Grouping

Guannan Lai, Yujie Li, Xiangkun Wang, Junbo Zhang, Tianrui Li, Xin Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4894-4904

Class Incremental Learning (CIL) aims to enable models to learn new classes sequentially while retaining knowledge of previous ones. Although current methods have alleviated catastrophic forgetting (CF), recent studies highlight that the performance of CIL models is highly sensitive to the order of class arrival, particularly when sequentially introduced classes exhibit high inter-class similarity. To address this critical yet understudied challenge of class order sensitivity, we first extend existing CIL frameworks through theoretical analysis, proving that grouping classes with lower pairwise similarity during incremental phases significantly improves model robustness to order variations. Building on this insight, we propose Graph-Driven Dynamic Similarity Grouping (GDDSG), a novel method that employs graph coloring algorithms to dynamically partition classes into similarity-constrained groups. Each group trains an isolated CIL sub-model and constructs meta-features for class group identification. Experimental results demonstrate that our method effectively addresses the issue of class order sensitivity while achieving optimal performance in both model accuracy and anti-forgetting capability. Our code is available at https://github.com/AIGNLAI/GDDSG.
****

Where the Devil Hides: Deepfake Detectors Can No Longer Be Trusted

Shuaiwei Yuan, Junyu Dong, Yuezun Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8764-8774

With the advancement of AI generative techniques, Deepfake faces have become incredibly realistic and nearly indistinguishable to the human eye. To counter this, Deepfake detectors have been developed as reliable tools for assessing face authenticity. These detectors are typically developed on Deep Neural Networks (DNNs) and trained using third-party datasets. However, this protocol raises a new security risk that can seriously undermine the trustfulness of Deepfake detectors: Once the third-party data providers insert poisoned (corrupted) data maliciously, Deepfake detectors trained on these datasets will be injected "backdoors" that cause abnormal behavior when presented with samples containing specific triggers. This is a practical concern, as third-party providers may distribute or sell these triggers to malicious users, allowing them to manipulate detector performance and escape accountability.This paper investigates this risk in depth and describes a solution to stealthily infect Deepfake detectors. Specifically, we develop a trigger generator, that can synthesize passcode-controlled, semantic-suppression, adaptive, and invisible trigger patterns, ensuring both the stealthiness and effectiveness of these triggers. Then we discuss two poisoning scenarios,

dirty-label poisoning and clean-label poisoning, to accomplish the injection of backdoors. Extensive experiments demonstrate the effectiveness, stealthiness, and practicality of our method compared to several baselines.

****

Synthetic-to-Real Self-supervised Robust Depth Estimation via Learning with Motion and Structure Priors

Weilong Yan, Ming Li, Haipeng Li, Shuwei Shao, Robby T. Tan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21880-21890

Self-supervised depth estimation from monocular cameras in diverse outdoor conditions, such as daytime, rain, and nighttime, is challenging due to the difficulty of learning universal representations and the severe lack of labeled real-world adverse data.Previous methods either rely on synthetic inputs and pseudo-depth labels or directly apply daytime strategies to adverse conditions, resulting in suboptimal results.In this paper, we present the first synthetic-to-real robust depth estimation framework, incorporating motion and structure priors to capture real-world knowledge effectively. In the synthetic adaptation, we transfer motion-structure knowledge inside cost volumes for better robust representation, using a frozen daytime model to train a depth estimator in synthetic adverse conditions.In the innovative real adaptation which targets to fix synthetic-real gaps, models trained earlier identify the weather-insensitive regions with a designed consistency-reweighting strategy to emphasize valid pseudo-labels.We further introduce a new regularization by gathering explicit depth distribution prior to constrain the model facing real-world data.Experiments show that our method outperforms the state-of-the-art across diverse conditions in multi-frame and single-frame settings. We achieve improvements of 7.5% in AbsRel and 4.3% in RMSE on average for nuScenes and Robotcar datasets (daytime, nighttime, rain). In zero-shot evaluation on DrivingStereo (rain, fog), our method generalizes better than previous ones. Our code will be released soon.

****

CaMuViD: Calibration-Free Multi-View Detection

Amir Etefaghi Daryani, M. Usman Maqbool Bhutta, Byron Hernandez, Henry Medeiros; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1220-1229

Multi-view object detection in crowded environments presents significant challenges, particularly for occlusion management across multiple camera views. This paper introduces a novel approach that extends conventional multi-view detection to operate directly within each camera's image space. Our method finds objects bounding boxes for images from various perspectives without resorting to a bird's eye view (BEV) representation. Thus, our approach removes the need for camera calibration by leveraging a learnable architecture that facilitates flexible transformations and improves feature fusion across perspectives to increase detection accuracy. Our model achieves Multi-Object Detection Accuracy (MODA) scores of 95.0% and 96.5% on the Wildtrack and MultiviewX datasets, respectively, significantly advancing the state of the art in multi-view detection. Furthermore, it demonstrates robust performance even without ground truth annotations, highlighting its resilience and practicality in real-world applications. These results emphasize the effectiveness of our calibration-free, multi-view object detector.

****

Prosody-Enhanced Acoustic Pre-training and Acoustic-Disentangled Prosody Adapting for Movie Dubbing

Zhedong Zhang, Liang Li, Chenggang Yan, Chunshan Liu, Anton van den Hengel, Yuankai Qi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 172-182

Movie dubbing describes the process of transforming a script into speech that aligns temporally and emotionally with a given movie clip while exemplifying the speaker's voice demonstrated in a short reference audio clip. This task demands the model bridge character performances and complicated prosody structures to build a high-quality video-synchronized dubbing track. The limited scale of movie dubbing datasets, along with the background noise inherent in audio data, hinder the acoustic modeling performance of trained models. To address these issues, we

propose an acoustic-prosody disentangled two-stage method to achieve high-quality dubbing generation with precise prosody alignment. First, we propose a prosody-enhanced acoustic pre-training to develop robust acoustic modeling capabilities. Then, we freeze the pre-trained acoustic system and design a disentangled framework to model prosodic text features and dubbing style while maintaining acoustic quality. Additionally, we incorporate an in-domain emotion analysis module to reduce the impact of visual domain shifts across different movies, thereby enhancing emotion-prosody alignment. Extensive experiments show that our method performs favorably against the state-of-the-art models on two primary benchmarks. The project is available at https://zzdoog.github.io/ProDubber/.

****

Hierarchical Knowledge Prompt Tuning for Multi-task Test-Time Adaptation

Qiang Zhang, Mengsheng Zhao, Jiawei Liu, Fanrui Zhang, Yongchao Xu, Zheng-Jun Zha; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30524-30533

Test-time adaptation using vision-language models (such as CLIP) to quickly adjust to distributional shifts of downstream tasks has shown great potential. Despite significant progress, existing methods are still limited to single-task test-time adaptation scenarios and have not effectively explored the issue of multi-task adaptation. To address this practical problem, we propose a novel Hierarchical Knowledge Prompt Tuning (HKPT) method, which achieves joint adaptation to multiple target domains by mining more comprehensive source domain discriminative knowledge and hierarchically modeling task-specific and task-shared knowledge. Specifically, HKPT constructs a CLIP prompt distillation framework that utilizes the broader source domain knowledge of large teacher CLIP to guide prompt tuning for lightweight student CLIP from multiple views during testing. Meanwhile, HKPT establishes task-specific dual dynamic knowledge graph to capture fine-grained contextual knowledge from continuous test data. To fully exploit the complementarity among multiple target tasks, HKPT employs an adaptive task grouping strategy for achieving inter-task knowledge sharing. Furthermore, HKPT can seamlessly transfer to basic single-task test-time adaptation scenarios while maintaining robust performance. Extensive experimental results in both multi-task and single-task test-time adaptation settings demonstrate that our HKPT significantly outperforms state-of-the-art methods.

****

Cross-Modal Interactive Perception Network with Mamba for Lung Tumor Segmentation in PET-CT Images

Jie Mei, Chenyu Lin, Yu Qiu, Yaonan Wang, Hui Zhang, Ziyang Wang, Dong Dai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15653-15662

Lung cancer is a leading cause of cancer-related deaths globally. PET-CT is crucial for imaging lung tumors, providing essential metabolic and anatomical information, while it faces challenges such as poor image quality, motion artifacts, and complex tumor morphology. Deep learning-based models are expected to address these problems, however, existing small-scale and private datasets limit significant performance improvements for these methods. Hence, we introduce a large-scale PET-CT lung tumor segmentation dataset, termed PCLT20K, which comprises 21,930 pairs of PET-CT images from 605 patients. Furthermore, we propose a cross-modal interactive perception network with Mamba (CIPA) for lung tumor segmentation in PET-CT images. Specifically, we design a channel-wise rectification module (CRM) that implements a channel state space block across multi-modal features to learn correlated representations and helps filter out modality-specific noise. A dynamic cross-modality interaction module (DCIM) is designed to effectively integrate position and context information, which employs PET images to learn regional position information and serves as a bridge to assist in modeling the relationships between local features of CT images. Extensive experiments on a comprehensive benchmark demonstrate the effectiveness of our CIPA compared to the current state-of-the-art segmentation methods. We hope our research can provide more exploration opportunities for medical image segmentation. The dataset and code are available at https://github.com/mj129/CIPA.

****
LaTexBlend: Scaling Multi-concept Customized Generation with Latent Textual Blending
Jian Jin, Zhenbo Yu, Yang Shen, Zhenyong Fu, Jian Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23585-23594
Customized text-to-image generation renders user-specified concepts into novel contexts based on textual prompts. Scaling the number of concepts in customized generation meets a broader demand for user creation, whereas existing methods face challenges with generation quality and computational efficiency. In this paper, we propose LaTexBlend, a novel framework for effectively and efficiently scaling multi-concept customized generation. The core idea of LaTexBlend is to represent single concepts and blend multiple concepts within a Latent Textual space, which is positioned after the text encoder and a linear projection. LaTexBlend customizes each concept individually, storing them in a concept bank with a compact representation of latent textual features that captures sufficient concept information to ensure high fidelity. At inference, concepts from the bank can be freely and seamlessly combined in the latent textual space, offering two key merits for multi-concept generation: 1) excellent scalability, and 2) significant reduction of denoising deviation, preserving coherent layouts. Extensive experiments demonstrate that LaTexBlend can flexibly integrate multiple customized concepts with harmonious structures and high subject fidelity, substantially outperforming baselines in both generation quality and computational efficiency. Our code will be publicly available.
****
DejaVid: Encoder-Agnostic Learned Temporal Matching for Video Classification
Darryl Ho, Samuel Madden; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24023-24032
In recent years, large transformer-based video encoder models have greatly advanced state-of-the-art performance on video classification tasks. However, these large models typically process videos by averaging embedding outputs from multiple clips over time to produce fixed-length representations. This approach fails to account for a variety of time-related features, such as variable video durations, chronological order of events, and temporal variance in feature significance. While methods for temporal modeling do exist, they often require significant architectural changes and expensive retraining, making them impractical for off-the-shelf, fine-tuned large encoders. To overcome these limitations, we propose DejaVid, an encoder-agnostic method that enhances model performance without the need for retraining or altering the architecture. Our framework converts a video into a variable-length temporal sequence of embeddings (TSE). A TSE naturally preserves temporal order and accommodates variable video durations. We then learn per-timestep, per-feature weights over the encoded TSE frames, allowing us to account for variations in feature importance over time. We introduce a new neural network architecture inspired by traditional time series alignment algorithms for this learning task. Our evaluation demonstrates that DejaVid substantially improves the performance of a state-of-the-art large encoder, achieving leading Top-1 accuracy of 77.2% on Something-Something V2, 89.1% on Kinetics-400, and 88.6% on HMDB51, while adding fewer than 1.8% additional learnable parameters and requiring less than 3 hours of training time. Our code is available at https://github.com/darrylho/DejaVid.
****
HVI: A New Color Space for Low-light Image Enhancement
Qingsen Yan, Yixu Feng, Cheng Zhang, Guansong Pang, Kangbiao Shi, Peng Wu, Wei Dong, Jinqiu Sun, Yanning Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5678-5687
Low-Light Image Enhancement (LLIE) is a crucial computer vision task that aims to restore detailed visual information from corrupted low-light images. Many existing LLIE methods are based on standard RGB (sRGB) space, which often produce color bias and brightness artifacts due to inherent high color sensitivity in sRGB. While converting the images using Hue, Saturation and Value (HSV) color space helps resolve the brightness issue, it introduces significant red and black nois

e artifacts. To address this issue, we propose a new color space for LLIE, namel y Horizontal/Vertical-Intensity (HVI), defined by polarized HS maps and learnabl e intensity. The former enforces small distances for red coordinates to remove t he red artifacts, while the latter compresses the low-light regions to remove th e black artifacts. To fully leverage the chromatic and intensity information, a novel Color and Intensity Decoupling Network (CIDNet) is further introduced to l earn accurate photometric mapping function under different lighting conditions i n the HVI space. Comprehensive results from benchmark and ablation experiments s how that the proposed HVI color space with CIDNet outperforms the state-of-the-a rt methods on 10 datasets. The code is available at https://github.com/Fediory/H VI-CIDNet.
****

DualPM: Dual Posed-Canonical Point Maps for 3D Shape and Pose Reconstruction
Ben Kaye, Tomas Jakab, Shangzhe Wu, Christian Ruprecht, Andrea Vedaldi; Proceedi ngs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6425-6435
The choice of data representation is a key factor in the success of deep learnin g in geometric tasks. For instance, DUSt3R has recently introduced the concept o f viewpoint- invariant point maps, generalizing depth prediction, and showing th at one can reduce all the key problems in the 3D reconstruction of static scenes to predicting such point maps. In this paper, we develop an analogous concept f or a very different problem, namely, the reconstruction of the 3D shape and pose of deformable objects. To this end, we introduce the Dual Point Maps (DualPM), where a pair of point maps is extracted from the same image, one associating pix els to their 3D locations on the object, and the other to a canonical version of the object at rest pose. We also extend point maps to amodal reconstruction, se eing through self-occlusions to obtain the complete shape of the object. We show that 3D reconstruction and 3D pose estimation reduce to the prediction of the D ualPMs. We demonstrate empirically that this representation is a good target for a deep network to predict; specifically, we consider modeling quadrupeds, showi ng that DualPMs can be trained purely on 3D synthetic data, consisting of one or two models per category, while generalizing very well to real images. With this , we improve by a large margin previous methods for the 3D analysis and reconstr uction of this type of objects.
****

SuperPC: A Single Diffusion Model for Point Cloud Completion, Upsampling, Denois ing, and Colorization
Yi Du, Zhipeng Zhao, Shaoshu Su, Sharath Golluri, Haoze Zheng, Runmao Yao, Chen Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVP R), 2025, pp. 16953-16964
Point cloud (PC) processing tasks--such as completion, upsampling, denoising, an d colorization--are crucial in applications like autonomous driving and 3D recon struction. Despite substantial advancements, prior approaches often address each of these tasks independently, with separate models focused on individual issues . However, this isolated approach fails to account for the fact that defects lik e incompleteness, low resolution, noise, and lack of color frequently coexist, w ith each defect influencing and correlating with the others.Simply applying thes e models sequentially can lead to error accumulation from each model, along with increased computational costs. To address these challenges, we introduce SuperP C, the first unified diffusion model capable of concurrently handling all four t asks. Our approach employs a three-level-conditioned diffusion framework, enhanc ed by a novel spatial-mix-fusion strategy, to leverage the correlations among th ese four defects for simultaneous, efficient processing.We show that SuperPC out performs the state-of-the-art specialized models as well as their combination on all four individual tasks.
****

One Diffusion to Generate Them All
Duong H. Le, Tuan Pham, Sangho Lee, Christopher Clark, Aniruddha Kembhavi, Steph an Mandt, Ranjay Krishna, Jiasen Lu; Proceedings of the Computer Vision and Patt ern Recognition Conference (CVPR), 2025, pp. 2671-2682

We introduce \texttt OneDiffusion - a single large-scale diffusion model design ed to tackle a wide range of image synthesis and understanding tasks. It can ge nerate images conditioned on text, depth, pose, layout, or semantic maps. It als o handles super-resolution, multi-view generation, instant personalization, and text-guided image editing. Additionally, the same model is also effective for im age understanding tasks such as depth estimation, pose estimation, open-vocab se gmentation and camera pose estimation, etc.Our unified model is trained with sim ple but effective recipe: we casting all tasks as modeling a sequence of frames where we inject different noise scales during training to each frame. During inf erence, we can set any frame to a clean image for tasks like conditional image g eneration, camera pose estimation, or generating paired depth and images from a caption. Experiment results shows our proposed method achieve competitive perfor mance on wide range of tasks. By eliminating the need for specialized architect ures or finetuning, OneDiffusion provides enhanced flexibility and scalability, reflecting the emerging trend towards general-purpose diffusion models in the fi eld.
****

Let's Verify and Reinforce Image Generation Step by Step
Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Ziyu Guo, Haoquan Zhang, Manyuan Zh ang, Jiaming Liu, Peng Gao, Hongsheng Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28662-28672
Chain-of-Thought (CoT) reasoning has been extensively explored in large models t o tackle complex understanding tasks. However, it still remains an open question whether such strategies can be applied to verifying and reinforcing image gener ation scenarios. In this paper, we provide the first comprehensive investigation in the potential of CoT reasoning to enhance autoregressive image generation. W e focus on three techniques: scaling test-time computation for verification, ali gning model preferences with Direct Preference Optimization (DPO), and integrati ng these techniques for complementary effects. Our results demonstrate that thes e approaches can be effectively adapted and combined to significantly improve im age generation performance. Furthermore, given the pivotal role of reward models in our findings, we propose the Potential Assessment Reward Model (PARM) specia lized for autoregressive image generation. PARM adaptively assesses each generat ion step through a potential assessment mechanism, merging the strengths of exis ting reward models. Using our investigated reasoning strategies, we enhance a ba seline model, Show-o, to achieve superior results, with a significant +24% impro vement on the GenEval benchmark, surpassing Stable Diffusion 3 by +15%. We hope our study provides unique insights and paves a new path for integrating CoT reas oning with autoregressive image generation. Code is released at https://github.c om/ZiyuGuo99/Image-Generation-CoT.
****

All-Optical Nonlinear Diffractive Deep Network for Ultrafast Image Denoising
Xiaoling Zhou, Zhemg Lee, Wei Ye, Rui Xie, Wenbo Zhang, Guanju Peng, Zongze Li, Shikun Zhang; Proceedings of the Computer Vision and Pattern Recognition Confere nce (CVPR), 2025, pp. 28221-28231
Image denoising poses a significant challenge in image processing, aiming to rem ove noise and artifacts from input images. However, current denoising algorithms implemented on electronic chips frequently encounter latency issues and demand substantial computational resources. In this paper, we introduce an all-optical Nonlinear Diffractive Denoising Deep Network (N3DNet) for image denoising at the speed of light. Initially, we incorporate an image encoding and pre-denoising m odule into the Diffractive Deep Neural Network and integrate a nonlinear activat ion function, termed the phase exponential linear function, after each diffracti ve layer, thereby boosting the network's nonlinear modeling and denoising capabi lities. Subsequently, we devise a new reinforcement learning algorithm called re gularization-assisted deep Q-network to optimize N3DNet. Finally, leveraging 3D printing techniques, we fabricate N3DNet using the trained parameters and constr uct a physical experimental system for real-world applications. A new benchmark dataset, termed MIDD, is constructed for mode image denoising, comprising 120K p airs of noisy/noise-free images captured from real fiber communication systems a

cross various transmission lengths. Through extensive simulation and real experi
ments, we validate that N3DNet outperforms both traditional and deep learning-ba
sed denoising approaches across various datasets. Remarkably, its processing spe
ed is nearly 3,800 times faster than electronic chip-based methods.
****

Maintaining Consistent Inter-Class Topology in Continual Test-Time Adaptation
Chenggong Ni, Fan Lyu, Jiayao Tan, Fuyuan Hu, Rui Yao, Tao Zhou; Proceedings of
the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15319-1
5328
This paper introduces Topological Consistency Adaptation (TCA), a novel approach
 to Continual Test-time Adaptation (CTTA) that addresses the challenges of domai
n shifts and error accumulation in testing scenarios. TCA ensures the stability
of inter-class relationships by enforcing a class topological consistency constr
aint, which minimizes the distortion of class centroids and preserves the topolo
gical structure during continuous adaptation. Additionally, we propose an intra-
class compactness loss to maintain compactness within classes, indirectly suppor
ting inter-class stability. To further enhance model adaptation, we introduce a
batch imbalance topology weighting mechanism that accounts for class distributio
n imbalances within each batch, optimizing centroid distances and stabilizing th
e inter-class topology. Experiments show that our method demonstrates improvemen
ts in handling continuous domain shifts, ensuring stable feature distributions a
nd boosting predictive performance.
****

UNOPose: Unseen Object Pose Estimation with an Unposed RGB-D Reference Image
Xingyu Liu, Gu Wang, Ruida Zhang, Chenyangguang Zhang, Federico Tombari, Xiangya
ng Ji; Proceedings of the Computer Vision and Pattern Recognition Conference (CV
PR), 2025, pp. 22023-22034
Unseen object pose estimation methods often rely on CAD models or multiple refer
ence views, making the onboarding stage costly. To simplify reference acquisitio
n, we aim to estimate the unseen object's pose through a single unposed RGB-D re
ference image. While previous works leverage reference images as pose anchors to
 limit the range of relative pose, our scenario presents significant challenges
since the relative transformation could vary across the entire SE(3) space. More
over, factors like occlusion, sensor noise, and extreme geometry could result in
 low viewpoint overlap. To address these challenges, we present a novel approach
 and benchmark, termed UNOPose, for UNseen One-reference-based object Pose estim
ation. Building upon a coarse-to-fine paradigm, UNOPose constructs an SE(3)-inv
ariant reference frame to standardize object representation despite pose and siz
e variations. To alleviate small overlap across viewpoints, we recalibrate the w
eight of each correspondence based on its predicted likelihood of being within t
he overlapping region. Evaluated on our proposed benchmark based on the BOP Chal
lenge, UNOPose demonstrates superior performance, significantly outperforming tr
aditional and learning-based methods in the one-reference setting and remaining
competitive with CAD-model-based methods. The code and dataset are available at
github.com/shanice-l/UNOPose.
****

CoSER: Towards Consistent Dense Multiview Text-to-Image Generator for 3D Creatio
n
Bonan Li, Zicheng Zhang, Xingyi Yang, Xinchao Wang; Proceedings of the Computer
Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2880-2890
Generating dense multiview images from text prompts is crucial for creating high
-fidelity 3D assets. Nevertheless, existing methods struggle with space-view cor
respondences, resulting in sparse and low-quality outputs. In this paper, we int
roduce CoSER, a novel consistent dense Multiview Text-to-Image Generator for Tex
t-to-3D, achieving both efficiency and quality by meticulously learning neighbor
-view coherence and further alleviating ambiguity through the swift traversal of
 all views. For achieving neighbor-view consistency, each viewpoint densely inte
racts with adjacent viewpoints to perceive the global spatial structure, and agg
regates information along motion paths explicitly defined by physical principles
 to refine details. To further enhance cross-view consistency and alleviate cont

ent drift, CoSER rapidly scan all views in spiral bidirectional manner to aware holistic information and then scores each point based on semantic material. Subsequently, we conduct weighted down-sampling along the spatial dimension based on scores, thereby facilitating prominent information fusion across all views with lightweight computation. Technically, the core module is built by integrating the attention mechanism with a selective state space model, exploiting the robust learning capabilities of the former and the low overhead of the latter. Extensive evaluation shows that CoSER is capable of producing dense, high-fidelity, content-consistent multiview images that can be flexibly integrated into various 3D generation models.

****

HybridMQA: Exploring Geometry-Texture Interactions for Colored Mesh Quality Assessment

Armin Shafiee Sarvestani, Sheyang Tang, Zhou Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21414-21424

Mesh quality assessment (MQA) models play a critical role in the design, optimization, and evaluation of mesh operation systems in a wide variety of applications. Current MQA models, whether model-based methods using topology-aware features or projection-based approaches working on rendered 2D projections, often fail to capture the intricate interactions between texture and 3D geometry. We introduce HybridMQA, a first-of-its-kind hybrid full-reference colored MQA framework that integrates model-based and projection-based approaches, capturing complex interactions between textural information and 3D structures for enriched quality representations. Our method employs graph learning to extract detailed 3D representations, which are then projected to 2D using a novel feature rendering process that precisely aligns them with colored projections. This enables the exploration of geometry-texture interactions via cross-attention, producing comprehensive mesh quality representations. Extensive experiments demonstrate HybridMQA's superior performance across diverse datasets, highlighting its ability to effectively leverage geometry-texture interactions for a thorough understanding of mesh quality. Our project website is available at https://arshafiee.github.io/hybridmqa/.

****

Generalized Gaussian Entropy Model for Point Cloud Attribute Compression with Dynamic Likelihood Intervals

Changhao Peng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11779-11788

Gaussian and Laplacian entropy models are proved effective in learned point cloud attribute compression, as they assist in arithmetic coding of latents. However, we demonstrate through experiments that there is still unutilized information in entropy parameters estimated by neural networks in current methods, which can be used for more accurate probability estimation. Thus we introduce generalized Gaussian entropy model, which controls the tail shape through shape parameter to more accurately estimate the probability of latents. Meanwhile, to the best of our knowledge, existing methods use fixed likelihood intervals for each integer during arithmetic coding, which limits model performance. We propose Mean Error Discriminator (MED) to determine whether the entropy parameter estimation is accurate and then dynamically adjust likelihood intervals. Experiments show that our method significantly improves rate-distortion (RD) performance on three VAE-based models for point cloud attribute compression, and our method can be applied to other compression tasks, such as image and video compression.

****

Self-Learning Hyperspectral and Multispectral Image Fusion via Adaptive Residual Guided Subspace Diffusion Model

Jian Zhu, He Wang, Yang Xu, Zebin Wu, Zhihui Wei; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17862-17871

Hyperspectral and multispectral image (HSI-MSI) fusion involves combining a low-resolution hyperspectral image (LR-HSI) with a high-resolution multispectral image (HR-MSI) to generate a high-resolution hyperspectral image (HR-HSI). Most deep learning-based methods for HSI-MSI fusion rely on large amounts of hyperspectr

al data for supervised training, which is often scarce in practical applications . In this paper, we propose a self-learning Adaptive Residual Guided Subspace Di ffusion Model (ARGS-Diff), which only utilizes the observed images without any e xtra training data. Specifically, as the LR-HSI contains spectral information an d the HR-MSI contains spatial information, we design two lightweight spectral an d spatial diffusion models to separately learn the spectral and spatial distribu tions from them. Then, we use these two models to reconstruct HR-HSI from two lo w-dimensional components, i.e, the spectral basis and the reduced coefficient, d uring the reverse diffusion process. Furthermore, we introduce an Adaptive Resid ual Guided Module (ARGM), which refines the two components through a residual gu ided function at each sampling step, thereby stabilizing the sampling process. E xtensive experimental results demonstrate that ARGS-Diff outperforms existing st ate-of-the-art methods in terms of both performance and computational efficiency in the field of HSI-MSI fusion.

****

SIR-DIFF: Sparse Image Sets Restoration with Multi-View Diffusion Model

Yucheng Mao, Boyang Wang, Nilesh Kulkarni, Jeong Joon Park; Proceedings of the C omputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21620-21630

The computer vision community has developed numerous techniques for digitally re storing true scene information from single-view degraded photographs, an importa nt yet extremely ill-posed task. In this work, we tackle image restoration from a different perspective by jointly denoising multiple photographs of the same sc ene. Our core hypothesis is that degraded images capturing a shared scene contai n complementary information that, when combined, better constrains the restorati on problem. To this end, we implement a powerful multi-view diffusion model that jointly generates uncorrupted views by extracting rich information from multi-v iew relationships. Our experiments show that our multi-view approach outperforms existing single-view image and even video-based methods on image deblurring and super-resolution tasks. Critically, our model is trained to output 3D consisten t images, making it a promising tool for applications requiring robust multi-vie w integration, such as 3D reconstruction or pose estimation.

****

StickMotion: Generating 3D Human Motions by Drawing a Stickman

Tao Wang, Zhihua Wu, Qiaozhi He, Jiaming Chu, Ling Qian, Yu Cheng, Junliang Xing , Jian Zhao, Lei Jin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12370-12379

Text-to-motion generation, which translates textual descriptions into human moti ons, has been challenging in accurately capturing detailed user-imagined motions from simple text inputs. This paper introduces StickMotion, an efficient diffus ion-based network designed for multi-condition scenarios, which generates desire d motions based on traditional text and our proposed stickman conditions for glo bal and local control of these motions, respectively. We address the challenges introduced by the user-friendly stickman from three perspectives: 1) Data genera tion. We develop an algorithm to generate hand-drawn stickmen automatically acro ss different dataset formats. 2) Multi-condition fusion. We propose a multi-cond ition module that integrates into the diffusion process and obtains outputs of a ll possible condition combinations, reducing computational complexity and enhanc ing StickMotion's performance compared to conventional approaches with the self- attention module. 3) Dynamic supervision. We empower StickMotion to make minor a djustments to the stickman's position within the output sequences, generating mo re natural movements through our proposed dynamic supervision strategy. Through quantitative experiments and user studies, sketching stickmen saves users about 51.5% of their time generating motions consistent with their imagination. Our co des, demos, and relevant data will be released to facilitate further research an d validation within the scientific community.

****

Reversible Decoupling Network for Single Image Reflection Removal

Hao Zhao, Mingjia Li, Qiming Hu, Xiaojie Guo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26430-26439

Recent deep-learning-based approaches to single-image reflection removal have sh

own promising advances, primarily for two reasons: 1) the utilization of recogni tion-pretrained features as inputs, and 2) the design of dual-stream interaction networks. However, according to the Information Bottleneck principle, high-leve l semantic clues tend to be compressed or discarded during layer-by-layer propag ation. Additionally, interactions in dual-stream networks follow a fixed pattern across different layers, limiting overall performance. To address these limitat ions, we propose a novel architecture called Reversible Decoupling Network (RDNe t), which employs a reversible encoder to secure valuable information while flex ibly decoupling transmission- and reflection-relevant features during the forwar d pass. Furthermore, we customize a transmission-rate-aware prompt generator to dynamically calibrate features, further boosting performance. Extensive experime nts demonstrate the superiority of RDNet over existing SOTA methods on five wide ly-adopted benchmark datasets. Our code will be made publicly available.

****

Hierarchical Features Matter: A Deep Exploration of Progressive Parameterization Method for Dataset Distillation

Xinhao Zhong, Hao Fang, Bin Chen, Xulin Gu, Meikang Qiu, Shuhan Qi, Shu-Tao Xia; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2 025, pp. 30462-30471

Dataset distillation is an emerging dataset reduction method, which condenses la rge-scale datasets while maintaining task accuracy. Current parameterization met hods achieve enhanced performance under extremely high compression ratio by opti mizing determined synthetic dataset in informative feature domain. However, they limit themselves to a fixed optimization space for distillation, neglecting the diverse guidance across different informative latent spaces. To overcome this l imitation, we propose a novel parameterization method dubbed Hierarchical Parame terization Distillation (H-PD), to systematically explore hierarchical feature w ithin provided feature space (e.g., layers within pre-trained generative adversa rial networks). We verify the correctness of our insights by applying the hierar chical optimization strategy on GAN-based parameterization method. In addition, we introduce a novel class-relevant feature distance metric to alleviate the com putational burden associated with synthetic dataset evaluation, bridging the gap between synthetic and original datasets. Experimental results demonstrate that the proposed H-PD achieves a significant performance improvement under various s ettings with equivalent time consumption, and even surpasses current generative distillation using diffusion models under extreme compression ratios IPC=1 and I PC=10.

****

Enduring, Efficient and Robust Trajectory Prediction Attack in Autonomous Drivin g via Optimization-Driven Multi-Frame Perturbation Framework

Yi Yu, Weizhen Han, Libing Wu, Bingyi Liu, Enshu Wang, Zhuangzhuang Zhang; Proce edings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, p p. 17229-17238

Trajectory prediction plays a crucial role in autonomous driving systems, and ex ploring its vulnerability has garnered widespread attention. However, existing t rajectory prediction attack methods often rely on single-point attacks to make e fficient perturbations. This limits their applications in real-world scenarios d ue to the transient nature of single-point attacks, their susceptibility to filt ration, and the uncertainty regarding the deployment environment. To address the se challenges, this paper proposes a novel LiDAR-induced attack framework to imp ose multi-frame attacks by optimization-driven adversarial location search, achi eving endurance, efficiency, and robustness. This framework strategically places objects near the adversarial vehicle to implement an attack and introduces thre e key innovations. First, successive state perturbations are generated using a m ulti-frame single-point attack strategy, effectively misleading trajectory predi ctions over extended time horizons. Second, we efficiently optimize adversarial objects' locations through three specialized loss functions to achieve desired p erturbations. Lastly, we improve robustness by treating the adversarial object a s a point without size constraints during the location search phase and reduce d ependence on both the specific attack point and the adversarial object's propert

ies. Extensive experiments confirm the superior performance and robustness of our framework.
****

GLASS: Guided Latent Slot Diffusion for Object-Centric Learning
Krishnakant Singh, Simone Schaub-Meyer, Stefan Roth; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28673-28683
Object-centric learning aims to decompose an input image into a set of meaningful object files (slots). These latent object representations enable a variety of downstream tasks. Yet, object-centric learning struggles on real-world datasets, which contain multiple objects of complex textures and shapes in natural everyday scenes. To address this, we introduce Guided Latent Slot Diffusion (GLASS), a novel slot attention model that learns in the space of generated images and uses semantic and instance guidance modules to learn better slot embeddings for various downstream tasks. Our experiments show that GLASS surpasses state-of-the-art slot attention methods by a wide margin on tasks such as (zero-shot) object discovery and conditional image generation for real-world scenes. Moreover, GLASS enables the first application of slot attention to compositional generation of complex, realistic scenes.
****

UNEM: UNrolled Generalized EM for Transductive Few-Shot Learning
Long Zhou, Fereshteh Shakeri, Aymen Sadraoui, Mounir Kaaniche, Jean-Christophe Pesquet, Ismail Ben Ayed; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9665-9675
Transductive few-shot learning has recently triggered wide attention in computer vision. Yet, current methods introduce key hyper-parameters, which control the pre-diction statistics of the test batches, such as the level of class balance, affecting performances significantly. Such hyper-parameters are empirically grid-searched over validation data, and their configurations may vary substantially with the target dataset and pre-training model, making such empirical searches both sub-optimal and computationally intractable. In this work, we advocate and introduce the unrolling paradigm, also referred to as "learning to optimize", in the context of few-shot learning, thereby learning efficiently and effectively a set of optimized hyperparameters. Specifically, we unroll a generalization of the ubiquitous Expectation-Maximization (EM) optimizer into a neural network architecture, mapping each of its iterates to a layer and learning a set of key hyper-parameters over validation data. Our unrolling approach covers various statistical feature distributions and pre-training paradigms, including recent foundational vision-language models and standard vision-only classifiers. We report comprehensive experiments, which cover a breadth of fine-grained downstream image classification tasks, showing significant gains brought by the proposed unrolled EM algorithm over iterative variants. The achieved improvements reach up to 10% and 7.5% on vision-only and vision-language benchmarks, respectively. The source code and learned parameters are available at https://github.com/ZhouLong0/UNEM-Transductive.
****

SASep: Saliency-Aware Structured Separation of Geometry and Feature for Open Set Learning on Point Clouds
Jinfeng Xu, Xianzhi Li, Yuan Tang, Xu Han, Qiao Yu, Yixue Hao, Long Hu, Min Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27295-27304
Recent advancements in deep learning have greatly enhanced 3D object recognition, but most models are limited to closed-set scenarios, unable to handle unknown samples in real-world applications. Open-set recognition (OSR) addresses this limitation by enabling models to both classify known classes and identify novel classes.However, current OSR methods rely on global features to differentiate known and unknown classes, treating the entire object uniformly and overlooking the varying semantic importance of its different parts.To address this gap, we propose Salience-Aware Structured Separation (SASep), which includes (i) a tunable semantic decomposition (TSD) module to semantically decompose objects into important and unimportant parts, (ii) a geometric synthesis strategy (GSS) to generate

pseudo-unknown objects by combining these unimportant parts, and (iii) a synth-a ided margin separation (SMS) module to enhance feature-level separation by expan ding the feature distributions between classes.Together, these components improv e both geometric and feature representations, enhancing the model's ability to e ffectively distinguish known and unknown classes.Experimental results show that SASep achieves superior performance in 3D OSR, outperforming existing state-of-t he-art methods. The codes are available at https://github.com/JinfengX/SASep.
****

Low-Biased General Annotated Dataset Generation

Dengyang Jiang, Haoyu Wang, Lei Zhang, Wei Wei, Guang Dai, Mengmeng Wang, Jingdo ng Wang, Yanning Zhang; Proceedings of the Computer Vision and Pattern Recogniti on Conference (CVPR), 2025, pp. 25113-25123

Pre-training backbone networks on a general annotated dataset (e.g., ImageNet) t hat comprises numerous manually collected images with category annotations has p roven to be indispensable for enhancing the generalization capacity of downstrea m visual tasks. However, those manually collected images often exhibit bias, whi ch is non-transferable across either categories or domains, thus causing the mod el's generalization capacity degeneration. To mitigate this problem, we present a low-biased general annotated dataset generation framework (lbGen). Instead of expensive manual collection, we aim at directly generating low-biased images wit h category annotations. To achieve this goal, we propose to leverage the advanta ge of a multimodal foundation model (e.g., CLIP), in terms of aligning images in a low-biased semantic space defined by language. Specifically, we develop a bi-level semantic alignment loss, which not only forces all generated images to be consistent with the semantic distribution of all categories belonging to the tar get dataset in an adversarial learning manner, but also requires each generated image to match the semantic description of its category name. In addition, we fu rther cast an existing image quality scoring model into a quality assurance loss to preserve the quality of the generated image. By leveraging these two loss fu nctions, we can obtain a low-biased image generation model by simply fine-tuning a pre-trained diffusion model using only all category names in the target datas et as input. Experimental results confirm that, compared with the manually label ed dataset or other synthetic datasets, the utilization of our generated low-bia sed dataset leads to stable generalization capacity enhancement of different bac kbone networks across various tasks, especially in tasks where the manually labe led samples are scarce.
****

G3Flow: Generative 3D Semantic Flow for Pose-aware and Generalizable Object Mani pulation

Tianxing Chen, Yao Mu, Zhixuan Liang, Zanxin Chen, Shijia Peng, Qiangyu Chen, Mi ngkun Xu, Ruizhen Hu, Hongyuan Zhang, Xuelong Li, Ping Luo; Proceedings of the C omputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1735-1744

Recent advances in imitation learning for 3D robotic manipulation have shown pro mising results with diffusion-based policies. However, achieving human-level dex terity requires seamless integration of geometric precision and semantic underst anding. We present G3Flow, a novel framework that constructs real-time semantic flow, a dynamic, object-centric 3D semantic representation by leveraging foundat ion models. Our approach uniquely combines 3D generative models for digital twin creation, vision foundation models for semantic feature extraction, and robust pose tracking for continuous semantic flow updates. This integration enables com plete semantic understanding even under occlusions while eliminating manual anno tation requirements. By incorporating semantic flow into diffusion policies, we demonstrate significant improvements in both terminal-constrained manipulation a nd cross-object generalization. Extensive experiments across five simulation tas ks show that G3Flow consistently outperforms existing approaches, achieving up t o 68.3% and 50.1% average success rates on terminal-constrained manipulation and cross-object generalization tasks respectively. Our results demonstrate the eff ectiveness of G3Flow in enhancing real-time dynamic semantic feature understandi ng for robotic manipulation policies.
****

Generative Hard Example Augmentation for Semantic Point Cloud Segmentation
Qi Zhang, Jibin Peng, Zhao Huang, Wei Feng, Di Lin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22205-22214
The recent progress in semantic point cloud segmentation is attributed to deep networks, which require a large amount of point cloud data for training. However, how to collect substantial point-wise annotations of the point clouds at affordable cost for the end-to-end network training still needs to be solved. In this paper, we propose Generative Hard Example Augmentation (GHEA) to achieve novel examples of point clouds, which enrich the data for training the segmentation network. Firstly, GHEA employs the generative network to embed the discrepancy between the point clouds into the latent space. From the latent space, we sample multiple discrepancies for reshaping a point cloud to various examples, contributing to the richness of the training data. Secondly, GHEA mixes the reshaped point clouds by respecting their segmentation errors. This mixup allows the reshaped point clouds, which are difficult to segment, to join as the challenging example for network training. We evaluate the effectiveness of GHEA, which helps the popular segmentation networks to improve the performances.
****

Toward Generalized Image Quality Assessment: Relaxing the Perfect Reference Quality Assumption
Du Chen, Tianhe Wu, Kede Ma, Lei Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12742-12752
Full-reference image quality assessment (FR-IQA) generally assumes that reference images are of perfect quality. However, this assumption is flawed due to the sensor and optical limitations of modern imaging systems. Moreover, recent generative enhancement methods are capable of producing images of higher quality than their original. All of these challenge the effectiveness and applicability of current FR-IQA models. To relax the assumption of perfect reference image quality, we build a large-scale IQA database, namely DiffIQA, containing approximately 180,000 images generated by a diffusion-based image enhancer with adjustable hyper-parameters. Each image is annotated by human subjects as either worse, similar, or better quality compared to its reference. Building on this, we present a generalized FR-IQA model, namely Adaptive FIdelity-Naturalness Evaluator (A-FINE), to accurately assess and adaptively combine the fidelity and naturalness of a test image. A-FINE aligns well with standard FR-IQA when the reference image is much more natural than the test image. We demonstrate by extensive experiments that A-FINE surpasses standard FR-IQA models on well-established IQA datasets and our newly created DiffIQA. To further validate A-FINE, we additionally construct a super-resolution IQA benchmark (SRIQA-Bench), encompassing test images derived from ten state-of-the-art SR methods with reliable human quality annotations. Tests on SRIQA-Bench re-affirm the advantages of A-FINE. The code and dataset are available at https://tianhewu.github.io/A-FINE-page.github.io/.
****

Explaining Domain Shifts in Language: Concept Erasing for Interpretable Image Classification
Zequn Zeng, Yudi Su, Jianqiao Sun, Tiansheng Wen, Hao Zhang, Zhengjue Wang, Bo Chen, Hongwei Liu, Jiawei Ma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9517-9526
Concept-based models can map black-box representations to human-understandable concepts, which makes the decision-making process more transparent and then allows users to understand the reason behind predictions. However, domain-specific concepts often impact the final predictions, which subsequently undermine the model generalization capabilities, and prevent the model from being used in high-stake applications. In this paper, we propose a novel Language-guided Concept-Erasing (LanCE) framework. In particular, we empirically demonstrate that pre-trained vision-language models (VLMs) can approximate distinct visual domain shifts via domain descriptors while prompting large Language Models (LLMs) can easily simulate a wide range of descriptors of unseen visual domains. Then, we introduce a novel plug-in domain descriptor orthogonality (DDO) regularizer to mitigate the impact of these domain-specific concepts on the final predictions. Notably, the

DDO regularizer is agnostic to the design of conceptbased models and we integrate it into several prevailing models. Through evaluation of domain generalization on four standard benchmarks and three newly introduced benchmarks, we demonstrate that DDO can significantly improve the out-of-distribution (OOD) generalization over the previous state-of-the-art concept-based models. Our code is available at https://github.com/joeyz0z/LanCE.

****

Hazy Low-Quality Satellite Video Restoration Via Learning Optimal Joint Degradation Patterns and Continuous-Scale Super-Resolution Reconstruction

Ning Ni, Libao Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12690-12699

Currently, the demand for higher video quality has grown significantly. However, satellite video has low resolution, complex motion, and weak textures. Haze interference further exacerbates the loss of motion information and texture details, hindering effective spatiotemporal feature fusion and fine-grained feature mining. This presents significant challenges for subsequent super-resolution (SR) reconstruction, especially at continuous scales. To address these problems, this paper models the double-degradation process of hazy low-quality satellite videos and proposes a novel network to learn the optimal joint degradation pattern (ODPNet) for continuous-scale SR of hazy satellite videos. First, we design a prior-based feature soft dehazing module to eliminate haze interference at the feature level. Second, we develop a spatiotemporal self-attention (SSA) to capture long-range feature dependencies, thereby achieving effective spatiotemporal feature fusion. Third, we devise a tri-branch cross-aggregation block (TCB) to enhance feature representations of weak textures in satellite videos by effectively aggregating contextual information. Finally, we propose a cross-scale feature Top-k selection Transformer (CFTST), which aims to adaptively select and aggregate cross-scale latent codes to learn feature representations of satellite videos at arbitrary resolutions, thus enabling continuous-scale SR. Experiments show that ODPNet outperforms existing methods and achieves a better balance between model parameters and performance.

****

Textured Gaussians for Enhanced 3D Scene Appearance Modeling

Brian Chao, Hung-Yu Tseng, Lorenzo Porzi, Chen Gao, Tuotuo Li, Qinbo Li, Ayush Saraf, Jia-Bin Huang, Johannes Kopf, Gordon Wetzstein, Changil Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8964-8974

3D Gaussian Splatting (3DGS) has recently emerged as a state-of-the-art 3D reconstruction and rendering technique due to its high-quality results and fast training and rendering time. However, pixels covered by the same Gaussian are always shaded in the same color up to a Gaussian falloff scaling factor. Furthermore, the finest geometric detail any individual Gaussian can represent is a simple ellipsoid. These properties of 3DGS greatly limit the expressivity of individual Gaussian primitives. To address these issues, we draw inspiration from texture and alpha mapping in traditional graphics and integrate it with 3DGS. Specifically, we propose a new generalized Gaussian appearance representation that augments each Gaussian with alpha (A), RGB, or RGBA texture maps to model spatially varying color and opacity across the extent of each Gaussian. As such, each Gaussian can represent a richer set of texture patterns and geometric structures, instead of just a single color and ellipsoid as in naive Gaussian Splatting. Surprisingly, we found that the expressivity of Gaussians can be greatly improved by using alpha-only texture maps, and further augmenting Gaussians with RGB texture maps achieves the highest expressivity. We validate our method on a wide variety of standard benchmark datasets and our own custom captures at both the object and scene levels. We demonstrate image quality improvements over existing methods while using a similar or lower number of Gaussians.

****

NeighborRetr: Balancing Hub Centrality in Cross-Modal Retrieval

Zengrong Lin, Zheng Wang, Tianwen Qian, Pan Mu, Sixian Chan, Cong Bai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9

263-9273
Cross-modal retrieval aims to bridge the semantic gap between different modaliti es, such as visual and textual data, enabling accurate retrieval across them. De spite significant advancements with models like CLIP that align cross-modal repr esentations, a persistent challenge remains: the hubness problem, where a small subset of samples (hubs) dominate as nearest neighbors, leading to biased repres entations and degraded retrieval accuracy. Existing methods often mitigate hubne ss through post-hoc normalization techniques, relying on prior data distribution s that may not be practical in real-world scenarios. In this paper, we directly mitigate hubness during training and introduce NeighborRetr, a novel method that effectively balances the learning of hubs and adaptively adjusts the relations of various kinds of neighbors. Our approach not only mitigates the hubness probl em but also enhances retrieval performance, achieving state-of-the-art results o n multiple cross-modal retrieval benchmarks. Furthermore, NeighborRetr demonstra tes robust generalization to new domains with substantial distribution shifts, h ighlighting its effectiveness in real-world applications.
****

ETAP: Event-based Tracking of Any Point
Friedhelm Hamann, Daniel Gehrig, Filbert Febryanto, Kostas Daniilidis, Guillermo Gallego; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27186-27196
Tracking any point (TAP) recently shifted the motion estimation paradigm from fo cusing on individual salient points with local templates to tracking arbitrary p oints with global image contexts. However, while research has mostly focused on driving the accuracy of models in nominal settings, addressing scenarios with di fficult lighting conditions and high-speed motions remains out of reach due to t he limitations of the sensor. This work addresses this challenge with the first event camera-based TAP method. It leverages the high temporal resolution and hig h dynamic range of event cameras for robust high-speed tracking, and the global contexts in TAP methods to handle asynchronous and sparse event measurements. We further extend the TAP framework to handle event feature variations induced by motion -- thereby addressing an open challenge in purely event-based tracking -- with a novel feature alignment-loss which ensures the learning of motion-robust features. Our method is trained with data from a new data generation pipeline a nd systematically ablated across all design decisions. Our method shows strong c ross-dataset generalization and performs 136% better on the average Jaccard metr ic than the baselines. Moreover, on an established feature tracking benchmark, i t achieves a 20% improvement over the previous best event-only method and even s urpasses the previous best events-and-frames method by 4.1%. Our code is availab le at https:// github.com/ tub-rip/ ETAP.
****

Beyond Sight: Towards Cognitive Alignment in LVLM via Enriched Visual Knowledge
Yaqi Zhao, Yuanyang Yin, Lin Li, Mingan Lin, Victor Shea-Jay Huang, Siwei Chen, Weipeng Chen, Baoqun Yin, Zenan Zhou, Wentao Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24950-24959
Does seeing always mean knowing? Large Vision-Language Models (LVLMs) integrate separately pre-trained vision and language components, often using CLIP-ViT as v ision backbone. However, these models frequently encounter a core issue of "cog nitive misalignment" between the vision encoder (VE) and the large language mode l (LLM). Specifically, the VE's representation of visual information may not ful ly align with LLM's cognitive framework, leading to a mismatch where visual feat ures exceed the language model's interpretive range. To address this, we investi gate how variations in VE representations influence LVLM comprehension, especial ly when the LLM faces VE-Unknown data--images whose ambiguous visual representat ions challenge the VE's interpretive precision. Accordingly, we construct a mult i-granularity landmark dataset and systematically examine the impact of VE-Known and VE-Unknown data on interpretive abilities. Our results show that VE-Unknown data limits LVLM's capacity for accurate understanding, while VE-Known data, ri ch in distinctive features, helps reduce cognitive misalignment. Building on the se insights, we propose Entity-Enhanced Cognitive Alignment (EECA), a method tha

t employs multi-granularity supervision to generate visually enriched, well-aligned tokens that not only integrate within the embedding space but also align with the LLM's cognitive framework. This alignment markedly enhances LVLM performance in landmark recognition. Our findings underscore the challenges posed by VE-Unknown data and highlight the essential role of cognitive alignment in advancing multimodal systems.

****

Global-Local Tree Search in VLMs for 3D Indoor Scene Generation
Wei Deng, Mengshi Qi, Huadong Ma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8975-8984
Large Vision-Language Models (VLMs), such as GPT-4, have achieved remarkable success across various fields. However, there are few studies on 3D indoor scene generation with VLMs. This paper considers this task as a planning problem subject to spatial and layout common sense constraints. To solve the problem with a VLM, we propose a new global-local tree search algorithm. Globally, the method places each object sequentially and explores multiple placements during each placement process, where the problem space is represented as a tree. To reduce the depth of the tree, we decompose the scene structure hierarchically, i.e. room level, region level, floor object level, and supported object level. The algorithm independently generates the floor objects in different regions and supported objects placed on different floor objects. Locally, we also decompose the sub-task, the placement of each object, into multiple steps. The algorithm searches the tree of problem space. To leverage the VLM model to produce positions of objects, we discretize the top-down view space as a dense grid and fill each cell with diverse emojis to make to cells distinct. We prompt the VLM with the emoji grid and the VLM produces a reasonable location for the object by describing the position with the name of emojis. The quantitative and qualitative experimental results illustrate our approach generates more plausible 3D scenes than state-of-the-art approaches. Our source code is available at https://github.com/dw-dengwei/TreeSearchGen .

****

Volumetric Surfaces: Representing Fuzzy Geometries with Layered Meshes
Stefano Esposito, Anpei Chen, Christian Reiser, Samuel Rota Bulò, Lorenzo Porzi, Katja Schwarz, Christian Richardt, Michael Zollhöfer, Peter Kontschieder, Andreas Geiger; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21370-21380
High-quality view synthesis relies on volume rendering, splatting, or surface rendering. While surface rendering is typically the fastest, it struggles to accurately model fuzzy geometry like hair. In turn, alpha-blending techniques excel at representing fuzzy materials but require an unbounded number of samples per ray (P1). Further overheads are induced by empty space skipping in volume rendering (P2) and sorting input primitives in splatting (P3). We present a novel representation for real-time view synthesis where the (P1) number of sampling locations is small and bounded, (P2) sampling locations are efficiently found via rasterization, and (P3) rendering is sorting-free. We achieve this by representing objects as semi-transparent multi-layer meshes rendered in a fixed order. First, we model surface layers as signed distance function (SDF) shells with optimal spacing learned during training. Then, we bake them as meshes and fit UV textures. Unlike single-surface methods, our multi-layer representation effectively models fuzzy objects. In contrast to volume and splatting-based methods, our approach enables real-time rendering on low-power laptops and smartphones.

****

Overcoming Shortcut Problem in VLM for Robust Out-of-Distribution Detection
Zhuo Xu, Xiang Xiang, Yifan Liang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15402-15412
Vision-language models (VLMs), such as CLIP, have shown remarkable capabilities in downstream tasks. However, the coupling of semantic information between the foreground and the background in images leads to significant shortcut issues that adversely affect out-of-distribution (OOD) detection abilities. When confronted with a background OOD sample, VLMs are prone to misidentifying it as in-distrib

ution (ID) data. In this paper, we analyze the OOD problem from the perspective of shortcuts in VLMs and propose OSPCoOp which includes background decoupling and mask-guided region regularization. We first decouple images into ID-relevant and ID-irrelevant regions and utilize the latter to generate a large number of augmented OOD background samples as pseudo-OOD supervision. We then use the masks from background decoupling to adjust the model's attention, minimizing its focus on ID-irrelevant regions. To assess the model's robustness against background interference, we introduce a new OOD evaluation dataset, ImageNet-Bg, which solely consists of background images with all ID-relevant regions removed. Our method demonstrates exceptional performance in few-shot scenarios, achieving strong results even in one-shot setting, and outperforms existing methods. The code and proposed ImageNet-Bg are available at https://github.com/HAIV-Lab/OSPCoOp_Imagenet-bg.

****

GFlowVLM: Enhancing Multi-step Reasoning in Vision-Language Models with Generative Flow Networks

Haoqiang Kang, Enna Sachdeva, Piyush Gupta, Sangjae Bae, Kwonjoon Lee; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3815-3825

Vision-Language Models (VLMs) have recently shown promising advancements in sequential decision-making tasks through task-specific fine-tuning. However, common fine-tuning methods, such as Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL) techniques like Proximal Policy Optimization (PPO), present notable limitations: SFT assumes Independent and Identically Distributed (IID) data, while PPO focuses on maximizing cumulative rewards. These limitations often restrict solution diversity and hinder generalization in multi-step reasoning tasks. To address these challenges, we introduce a novel framework, GFlowVLM, a framework that fine-tune VLMs using Generative Flow Networks (GFlowNets) to promote generation of diverse solutions for complex reasoning tasks. GFlowVLM models the environment as a non-Markovian decision process, allowing it to capture long-term dependencies essential for real-world applications. It takes observations and task descriptions as inputs to prompt chain-of-thought (CoT) reasoning which subsequently guides action selection. We use task based rewards to fine-tune VLM with GFlowNets. This approach enables VLMs to outperform prior fine-tuning methods, including SFT and RL. Empirical results demonstrate the effectiveness of GFlowVLM on complex tasks such as card games (NumberLine, BlackJack) and embodied planning tasks (ALFWorld), showing enhanced training efficiency, solution diversity, and stronger generalization capabilities across both in-distribution and out-of-distribution scenarios. Project page is available at \href https://mk322.github.io/gflowvlm/ https://mk322.github.io/gflowvlm/ .

****

STEPS: Sequential Probability Tensor Estimation for Text-to-Image Hard Prompt Search

Yuning Qiu, Andong Wang, Chao Li, Haonan Huang, Guoxu Zhou, Qibin Zhao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28640-28650

Recent text-to-image (T2I) diffusion models have demonstrated remarkable capabilities in visual synthesis, yet their performance heavily relies on the quality of input prompts. However, optimizing discrete prompts remains challenging because the discrete nature of tokens prevents the direct application of the gradient descent method and the vast search space of possible token combinations. As a result, existing approaches either suffer from quantization errors when employing continuous optimization techniques or become trapped in local optima due to coordinate-wise greedy search. In this paper, we propose STEPS, a novel Sequential probability Tensor Estimation approach for hard Prompt Search. Our method reformulates discrete prompt optimization as a sequential probability tensor estimation problem, leveraging the inherent low-rank characteristics to address the curse of dimensionality. To further improve the computational efficiency, we develop a memory-bounded sampling approach that shrinks the prompt space without the iteration step dependency while preserving sequential optimization dynamics. Extensi

ve experiments on various public datasets demonstrate that our method consistent
ly outperforms existing approaches in T2I generation, cross-model prompt transfe
rability, and harmful prompt optimization, validating the effectiveness of the p
roposed framework.
****

RoboSpatial: Teaching Spatial Understanding to 2D and 3D Vision-Language Models
for Robotics
Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, Stan Birch
field; Proceedings of the Computer Vision and Pattern Recognition Conference (CV
PR), 2025, pp. 15768-15780
Spatial understanding is a crucial capability that enables robots to perceive th
eir surroundings, reason about their environment, and interact with it meaningfu
lly. In modern robotics, these capabilities are increasingly provided by vision-
language models. However, these models face significant challenges in spatial re
asoning tasks, as their training data are based on general-purpose image dataset
s that often lack sophisticated spatial understanding. For example, datasets fre
quently do not capture reference frame comprehension, yet effective spatial reas
oning requires understanding whether to reason from ego-, world-, or object-cent
ric perspectives. To address this issue, we introduce RoboSpatial, a large-scale
 dataset for spatial understanding in robotics. It consists of real indoor and
tabletop scenes, captured as 3D scans and egocentric images, and annotated with
rich spatial information relevant to robotics. The dataset includes 1M images, 5
k 3D scans, and 3M annotated spatial relationships, and the pairing of 2D egocen
tric images with 3D scans makes it both 2D- and 3D- ready. Our experiments show
that models trained with RoboSpatial outperform baselines on downstream tasks su
ch as spatial affordance prediction, spatial relationship prediction, and roboti
cs manipulation.
****

VIRES: Video Instance Repainting via Sketch and Text Guided Generation
Shuchen Weng, Haojie Zheng, Peixuan Zhang, Yuchen Hong, Han Jiang, Si Li, Boxin
Shi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR
), 2025, pp. 28416-28425
We introduce VIRES, a video instance repainting method with sketch and text guid
ance, enabling video instance repainting, replacement, generation, and removal.
Existing approaches struggle with temporal consistency and accurate alignment wi
th the provided sketch sequence. VIRES leverages the generative priors of text-t
o-video models to maintain temporal consistency and produce visually pleasing re
sults. We propose the Sequential ControlNet with the standardized self-scaling,
which effectively extracts structure layouts and adaptively captures high-contra
st sketch details. We further augment the diffusion transformer backbone with th
e sketch attention to interpret and inject fine-grained sketch semantics. A sket
ch-aware encoder ensures that repainted results are aligned with the provided sk
etch sequence. Additionally, we contribute the VireSet, a dataset with detailed
annotations tailored for training and evaluating video instance editing methods.
 Experimental results demonstrate the effectiveness of VIRES, which outperforms
state-of-the-art methods in visual quality, temporal consistency, condition alig
nment, and human ratings. The code, dataset and pretrained models are available
at: https://hjzheng.net/projects/VIRES.
****

MAP: Unleashing Hybrid Mamba-Transformer Vision Backbone's Potential with Masked
 Autoregressive Pretraining
Yunze Liu, Li Yi; Proceedings of the Computer Vision and Pattern Recognition Con
ference (CVPR), 2025, pp. 9676-9685
Hybrid Mamba-Transformer networks have recently garnered broad attention. These
networks can leverage the scalability of Transformers while capitalizing on Mamb
a's strengths in long-context modeling and computational efficiency. However, th
e challenge of effectively pretraining such hybrid networks remains an open ques
tion. Existing methods, such as Masked Autoencoders (MAE) or autoregressive (AR)
 pretraining, primarily focus on single-type network architectures. In contrast,
 pretraining strategies for hybrid architectures must be effective for both Mamb

a and Transformer components. Based on this, we propose Masked Autoregressive Pr
etraining (MAP) to pretrain a hybrid Mamba-Transformer vision backbone network.
This strategy combines the strengths of both MAE and Autoregressive pretraining,
 improving the performance of Mamba and Transformer modules within a unified par
adigm. Experimental results show that the hybrid Mamba-Transformer vision backbo
ne network pretrained with MAP significantly outperforms other pretraining strat
egies, achieving state-of-the-art performance. We validate the method's effectiv
eness on both 2D and 3D datasets and provide detailed ablation studies to suppor
t the design choices for each component.
****

Segment Any-Quality Images with Generative Latent Space Enhancement
Guangqian Guo, Yong Guo, Xuehui Yu, Wenbo Li, Yaoxing Wang, Shan Gao; Proceeding
s of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23
66-2376
Despite their success, Segment Anything Models (SAMs) experience significant per
formance drops on severely degraded, low-quality images, limiting their effectiv
eness in real-world scenarios. To address this, we propose GleSAM, which utilize
s Generative Latent space Enhancement to boost robustness on low-quality images,
 thus enabling generalization across various image qualities. Specifically, we
adapt the concept of latent diffusion to SAM-based segmentation frameworks and p
erform the generative diffusion process in the latent space of SAM to reconstruc
t high-quality representation, thereby improving segmentation. Additionally, we
introduce two techniques to improve compatibility between the pre-trained diffus
ion model and the segmentation framework. Our method can be applied to pre-train
ed SAM and SAM2 with only minimal additional learnable parameters, allowing for
efficient optimization. We also construct the LQSeg dataset with a greater diver
sity of degradation types and levels for training and evaluating the model. Exte
nsive experiments demonstrate that GleSAM significantly improves segmentation ro
bustness on complex degradations while maintaining generalization to clear image
s. Furthermore, GleSAM also performs well on unseen degradations, underscoring t
he versatility of our approach and dataset. Codes and datasets will be available
 soon.
****

BG-Triangle: Bezier Gaussian Triangle for 3D Vectorization and Rendering
Minye Wu, Haizhao Dai, Kaixin Yao, Tinne Tuytelaars, Jingyi Yu; Proceedings of t
he Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16197-16
207
Differentiable rendering enables efficient optimization by allowing gradients to
 be computed through the rendering process, facilitating 3D reconstruction, inve
rse rendering and neural scene representation learning. To ensure differentiabil
ity, existing solutions approximate or re-formulate traditional rendering operat
ions using smooth, probabilistic proxies such as volumes or Gaussian primitives.
 Consequently, they struggle to preserve sharp edges due to the lack of explicit
 boundary definitions. We present a novel hybrid representation, Bezier Gaussian
 Triangle (BG-Triangle), that combines Bezier triangle-based vector graphics pri
mitives with Gaussian-based probabilistic models, to maintain accurate shape mod
eling while conducting resolution-independent differentiable rendering. We prese
nt a robust and effective discontinuity-aware rendering technique to reduce unce
rtainties at object boundaries. We also employ an adaptive densification and pru
ning scheme for efficient training while reliably handling level-of-detail (LoD)
 variations. Experiments show that BG-Triangle achieves comparable rendering qua
lity as 3DGS but with superior boundary preservation. More importantly, BG-Trian
gle uses a much smaller number of primitives than its alternatives, showcasing t
he benefits of vectorized graphics primitives and the potential to bridge the ga
p between classic and emerging representations.
****

MIMO: Controllable Character Video Synthesis with Spatial Decomposed Modeling
Yifang Men, Yuan Yao, Miaomiao Cui, Liefeng Bo; Proceedings of the Computer Visi
on and Pattern Recognition Conference (CVPR), 2025, pp. 21181-21191
Character video synthesis aims to produce realistic videos of animatable charact

ers within lifelike scenes. As a fundamental problem in the computer vision and graphics community, 3D works typically require multi-view captures for per-case training, which severely limits their applicability of modeling arbitrary characters in a short time. Recent 2D methods break this limitation via pre-trained diffusion models, but they struggle for flexible controls, pose generality and scene interaction. To this end, we propose MIMO, a novel framework which can not only synthesize realistic character videos with controllable attributes (i.e., character, motion and scene) provided by simple user inputs, but also simultaneously achieve advanced scalability to arbitrary characters, generality to novel 3D motions, and applicability to interactive real-world scenes in a unified framework. The core idea is to encode the 2D video to compact spatial codes, considering the inherent 3D nature of video occurrence. Concretely, we lift the 2D frame pixels into 3D using monocular depth estimators, and decompose the video clip into three spatial components (i.e., main human, underlying scene, and floating occlusion) in hierarchical layers based on the 3D depth. These components are further encoded to canonical identity code, structured motion code and full scene code, which are utilized as control signals of the synthesis process. The design of spatial decomposed modeling enables flexible user control, complex motion expression, as well as 3D-aware synthesis for scene interactions. Experimental results show that the proposed method outperforms prior works by a large margin in character animation synthesis and is effective in providing a high degree of controllability (i.e., arbitrary characters, novel 3D motions, interactive scenes), thus enabling brand-new editing tasks (e.g., video character replacement).
****

TKG-DM: Training-free Chroma Key Content Generation Diffusion Model
Ryugo Morita, Stanislav Frolov, Brian Bernhard Moser, Takahiro Shirakawa, Ko Watanabe, Andreas Dengel, Jinjia Zhou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13031-13040
Diffusion models have enabled the generation of high-quality images with a strong focus on realism and textual fidelity. Yet, large-scale text-to-image models, such as Stable Diffusion, struggle to generate images where foreground objects are placed over a chroma key background, limiting their ability to separate foreground and background elements without fine-tuning. To address this limitation, we present a novel Training-Free Chroma Key Content Generation Diffusion Model (TKG-DM), which optimizes the initial random noise to produce images with foreground objects on a specifiable color background. Our proposed method is the first to explore the manipulation of the color aspects in initial noise for controlled background generation, enabling precise separation of foreground and background without fine-tuning. Extensive experiments demonstrate that our training-free method outperforms existing methods in both qualitative and quantitative evaluations, matching or surpassing fine-tuned models. Finally, we successfully extend it to other tasks (e.g., consistency models and text-to-video), highlighting its transformative potential across various generative applications where independent control of foreground and background is crucial.
****

Lift3D Policy: Lifting 2D Foundation Models for Robust 3D Robotic Manipulation
Yueru Jia, Jiaming Liu, Sixiang Chen, Chenyang Gu, Zhilve Wang, Longzan Luo, Xiaoqi Li, Pengwei Wang, Zhongyuan Wang, Renrui Zhang, Shanghang Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17347-17358
3D geometric information is essential for manipulation tasks, as robots need to perceive the 3D environment, reason about spatial relationships, and interact with intricate spatial configurations. Recent research has increasingly focused on the explicit extraction of 3D features, while still facing challenges such as the lack of large-scale robotic 3D data and the potential loss of spatial geometry. To address these limitations, we propose the Lift3D framework, which progressively enhances 2D foundation models with implicit and explicit 3D robotic representations to construct a robust 3D manipulation policy. Specifically, we first design a task-aware masked autoencoder that masks task-relevant affordance patches and reconstructs depth information, enhancing the 2D foundation model's implic

it 3D robotic representation. After self-supervised fine-tuning, we introduce a 2D model-lifting strategy that establishes a positional mapping between the input 3D points and the positional embeddings of the 2D model. Based on the mapping, Lift3D utilizes the 2D foundation model to directly encode point cloud data, leveraging large-scale pretrained knowledge to construct explicit 3D robotic representations while minimizing spatial information loss. In experiments, Lift3D consistently outperforms previous state-of-the-art methods across several simulation benchmarks and real-world scenarios.
****

Multi-View Pose-Agnostic Change Localization with Zero Labels
Chamuditha Jayanga Galappaththige, Jason Lai, Lloyd Windrim, Donald Dansereau, Niko Sunderhauf, Dimity Miller; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11600-11610
Autonomous agents often require accurate methods for detecting and localizing changes in their environment, particularly when observations are captured from unconstrained and inconsistent viewpoints. We propose a novel label-free, pose-agnostic change detection method that integrates information from multiple viewpoints to construct a change-aware 3D Gaussian Splatting (3DGS) representation of the scene. With as few as 5 images of the post-change scene, our approach can learn an additional change channel in a 3DGS and produce change masks that outperform single-view techniques. Our change-aware 3D scene representation additionally enables the generation of accurate change masks for unseen viewpoints. Experimental results demonstrate state-of-the-art performance in complex multi-object scenes, achieving a 1.7x and 1.5x improvement in Mean Intersection Over Union and F1 score respectively over other baselines. We also contribute a new real-world dataset to benchmark change detection in diverse challenging scenes in the presence of lighting variations.
****

From Sparse to Dense: Camera Relocalization with Scene-Specific Detector from Feature Gaussian Splatting
Zhiwei Huang, Hailin Yu, Yichun Shentu, Jin Yuan, Guofeng Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27059-27069
This paper presents a novel camera relocalization method, STDLoc, which leverages Feature Gaussian as scene representation. STDLoc is a full relocalization pipeline that can achieve accurate relocalization without relying on any pose prior. Unlike previous coarse-to-fine localization methods that require image retrieval first and then feature matching, we propose a novel sparse-to-dense localization paradigm. Based on this scene representation, we introduce a novel matching-oriented Gaussian sampling strategy and a scene-specific detector to achieve efficient and robust initial pose estimation. Furthermore, based on the initial localization results, we align the query feature map to the Gaussian feature field by dense feature matching to enable accurate localization. The experiments on indoor and outdoor datasets show that STDLoc outperforms current state-of-the-art localization methods in terms of localization accuracy and recall. Our code is available on the project website: https://zju3dv.github.io/STDLoc.
****

Accelerating Diffusion Transformer via Increment-Calibrated Caching with Channel-Aware Singular Value Decomposition
Zhiyuan Chen, Keyi Li, Yifan Jia, Le Ye, Yufei Ma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18011-18020
Diffusion transformer (DiT) models have achieved remarkable success in image generation, thanks for their exceptional generative capabilities and scalability. Nonetheless, the iterative nature of diffusion models (DMs) results in high computation complexity, posing challenges for deployment. Although existing cache-based acceleration methods try to utilize the inherent temporal similarity to skip redundant computations of DiT, the lack of correction may induce potential quality degradation. In this paper, we propose increment-calibrated caching, a training-free method for DiT acceleration, where the calibration parameters are generated from the pre-trained model itself with low-rank approximation. To deal with

the possible correction failure arising from outlier activations, we introduce c
hannel-aware Singular Value Decomposition (SVD), which further strengthens the c
alibration effect. Experimental results show that our method always achieve bett
er performance than existing naive caching methods with a similar computation re
source budget. When compared with 35-step DDIM, our method eliminates more than
45% computation and improves IS by 12 at the cost of less than 0.06 FID increase
.
****

CityWalker: Learning Embodied Urban Navigation from Web-Scale Videos
Xinhao Liu, Jintong Li, Yicheng Jiang, Niranjan Sujay, Zhicheng Yang, Juexiao Zh
ang, John Abanes, Jing Zhang, Chen Feng; Proceedings of the Computer Vision and
Pattern Recognition Conference (CVPR), 2025, pp. 6875-6885
Navigating dynamic urban environments presents significant challenges for embodi
ed agents, requiring advanced spatial reasoning and adherence to common-sense no
rms. Despite progress, existing visual navigation methods struggle in map-free o
r off-street settings, limiting the deployment of autonomous agents like last-mi
le delivery robots. To overcome these obstacles, we propose a scalable, data-dri
ven approach for human-like urban navigation by training agents on thousands of
hours of in-the-wild city walking and driving videos sourced from the web. We in
troduce a simple and scalable data processing pipeline that extracts action supe
rvision from these videos, enabling large-scale imitation learning without costl
y annotations. Our model learns sophisticated navigation policies to handle dive
rse challenges and critical scenarios. Experimental results show that training o
n large-scale, diverse datasets significantly enhances navigation performance, s
urpassing current methods. This work shows the potential of using abundant onlin
e video data to develop robust navigation policies for embodied agents in dynami
c urban settings.
****

A Simple yet Effective Layout Token in Large Language Models for Document Unders
tanding
Zhaoqing Zhu, Chuwei Luo, Zirui Shao, Feiyu Gao, Hangdi Xing, Qi Zheng, Ji Zhang
; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR),
2025, pp. 14472-14482
Recent methods that integrate spatial layouts with text for document understandi
ng in large language models (LLMs) have shown promising results. A commonly used
 method is to represent layout information as text tokens and interleave them wi
th text content as inputs to the LLMs. However, such a method still demonstrates
 limitations, as it requires additional position IDs for tokens that are used to
 represent layout information. Due to the constraint on max position IDs, assign
ing them to layout information reduces those available for text content, reducin
g the capacity for the model to learn from the text during training, while also
introducing a large number of potentially untrained position IDs during long-con
text inference, which can hinder performance on document understanding tasks. To
 address these issues, we propose LayTokenLLM, a simple yet effective method for
 document understanding. LayTokenLLM represents layout information as a single t
oken per text segment and uses a specialized positional encoding scheme. It shar
es position IDs between text and layout tokens, eliminating the need for additio
nal position IDs. This design maintains the model's capacity to learn from text
while mitigating long-context issues during inference. Furthermore, a novel pre-
training objective called Next Interleaved Text and Layout Token Prediction (NTL
P) is devised to enhance cross-modality learning between text and layout tokens.
 Extensive experiments show that LayTokenLLM outperforms existing layout-integra
ted LLMs and MLLMs of similar scales on multi-page document understanding tasks,
 as well as most single-page tasks.
****

Reconstruction vs. Generation: Taming Optimization Dilemma in Latent Diffusion M
odels
Jingfeng Yao, Bin Yang, Xinggang Wang; Proceedings of the Computer Vision and Pa
ttern Recognition Conference (CVPR), 2025, pp. 15703-15712
Latent diffusion models with Transformer architectures excel at generating high-

fidelity images. However, recent studies reveal an optimization dilemma in this two-stage design: while increasing the per-token feature dimension in visual tok enizers improves reconstruction quality, it requires substantially larger diffus ion models and more training iterations to achieve comparable generation perform ance. Consequently, existing systems often settle for sub-optimal solutions, eit her producing visual artifacts due to information loss within tokenizers or fail ing to converge fully due to expensive computation costs. We argue that this dil emma stems from the inherent difficulty in learning unconstrained high-dimension al latent spaces. To address this, we propose aligning the latent space with pre -trained vision foundation models when training the visual tokenizers. Our propo sed VA-VAE (Vision foundation model Aligned Variational AutoEncoder) significant ly expands the reconstruction-generation frontier of latent diffusion models, en abling faster convergence of Diffusion Transformers (DiT) in high-dimensional la tent spaces. To exploit the full potential of VA-VAE, we build an enhanced DiT b aseline with improved training strategies and architecture designs, termed Light ningDiT. The integrated system achieves state-of-the-art (SOTA) performance on I mageNet 256 generation with an FID score of 1.35 while demonstrating remarkable training efficiency by reaching an FID score of 2.11 in just 64 epochs -- repres enting an over 21 times convergence speedup compared to the original DiT. Models and codes are available at https://github.com/hustvl/LightningDiT.

****

StableAnimator: High-Quality Identity-Preserving Human Image Animation

Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, Zuxuan Wu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 20 25, pp. 21096-21106

Current diffusion models for human image animation struggle to ensure identity (ID) consistency. This paper presents StableAnimator, the first end-to-end ID-pre serving video diffusion framework, which synthesizes high-quality videos without any post-processing, conditioned on a reference image and a sequence of poses. Building upon a video diffusion model, StableAnimator contains carefully designe d modules for both training and inference striving for identity consistency. In particular, StableAnimator begins by computing image and face embeddings with of f-the-shelf extractors, respectively and face embeddings are further refined by interacting with image embeddings using a global content-aware Face Encoder. The n, StableAnimator introduces a novel distribution-aware ID Adapter that prevents interference caused by temporal layers while preserving ID via alignment. Durin g inference, we propose a novel Hamilton-Jacobi-Bellman (HJB) equation-based opt imization to further enhance the face quality. We demonstrate that solving the H JB equation can be integrated into the diffusion denoising process, and the resu lting solution constrains the denoising path and thus benefits ID preservation. Experiments on multiple benchmarks show the effectiveness of StableAnimator both qualitatively and quantitatively.

****

Learning Visual Composition through Improved Semantic Guidance

Austin Stone, Hagen Soltau, Robert Geirhos, Xi Yi, Ye Xia, Bingyi Cao, Kaifeng C hen, Abhijit Ogale, Jonathon Shlens; Proceedings of the Computer Vision and Patt ern Recognition Conference (CVPR), 2025, pp. 3740-3750

Visual imagery does not consist of solitary objects, but in-stead reflects the c omposition of a multitude of fluid con-cepts. While there have been great advanc es in visual repre-sentation learning, such advances have focused on buildingbet ter representations for a small number of discrete objectsbereft of an understan ding of how these objects are inter-acting. One can observe this limitation in representationslearned through captions or contrastive learning - wherethe learn ed model treats an image essentially as a bag ofwords. Several works have attem pted to address this lim-itation through the development of bespoke archit ectures.In this work, we focus on simple and scalable approaches.In particular, we demonstrate that by improving weakly la-beled data, i.e. captions, we can vas tly improve the perfor-mance of standard contrastive learning approaches. Previ-ous CLIP models achieved near chance rate on challengingtasks probing compositio nal learning. However, our sim-ple approach boosts performance of CLIP substa

ntially andachieves state of the art results on compositional bench-marks such as ARO and SugarCrepe. Furthermore, weshowcase our results on a relatively new captioning bench-mark derived from DOCCI. We demonstrate through a se-ries of ablations that a standard CLIP model trained withenhanced data may demonstrate impressive performance onimage retrieval tasks.
****

OODD: Test-time Out-of-Distribution Detection with Dynamic Dictionary
Yifeng Yang, Lin Zhu, Zewen Sun, Hengyu Liu, Qinying Gu, Nanyang Ye; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30630-30639
Out-of-distribution (OOD) detection remains challenging for deep learning models, particularly when test-time OOD samples differ significantly from training outliers. We propose OODD, a novel test-time OOD detection method that dynamically maintains and updates an OOD dictionary without fine-tuning. Our approach leverages a priority queue-based dictionary that accumulates representative OOD features during testing, combined with an informative inlier sampling strategy for in-distribution (ID) samples. To ensure stable performance during early testing, we propose a dual OOD stabilization mechanism that leverages strategically generated outliers derived from ID data. To our best knowledge, extensive experiments on the OpenOOD benchmark demonstrate that OODD significantly outperforms existing methods, achieving a 26.0% improvement in FPR95 on CIFAR-100 Far OOD detection compared to the state-of-the-art approach. Furthermore, we present an optimized variant of the KNN-based OOD detection framework that achieves a 3x speedup while maintaining detection performance.
****

MEAT: Multiview Diffusion Model for Human Generation on Megapixels with Mesh Attention
Yuhan Wang, Fangzhou Hong, Shuai Yang, Liming Jiang, Wayne Wu, Chen Change Loy; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11297-11306
Multiview diffusion models have shown considerable success in image-to-3D generation for general objects. However, when applied to human data, existing methods have yet to deliver promising results, largely due to the challenges of scaling multiview attention to higher resolutions. In this paper, we explore human multiview diffusion models at the megapixel level and introduce a solution called "mesh attention" to enable training at 1024x1024 resolution. Using a clothed human mesh as a central coarse geometric representation, the proposed mesh attention leverages rasterization and projection to establish direct cross-view coordinate correspondences. This approach significantly reduces the complexity of multiview attention while maintaining cross-view consistency. Building on this foundation, we devise a mesh attention block and combine it with keypoint conditioning to create our human-specific multiview diffusion model, MEAT. In addition, we present valuable insights into applying multiview human motion videos for diffusion training, addressing the longstanding issue of data scarcity. Extensive experiments show that MEAT effectively generates dense, consistent multiview human images at the megapixel level, outperforming existing multiview diffusion methods. Code is available at https://johann.wang/MEAT/.
****

Free Lunch Enhancements for Multi-modal Crowd Counting
Haoliang Meng, Xiaopeng Hong, Zhengqin Lai, Miao Shang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14013-14023
This paper addresses multi-modal crowd counting with a novel `free lunch' training enhancement strategy that requires no additional data, parameters, or increased inference complexity. First, we introduce a cross-modal alignment technique as a plug-in post-processing step for the pre-trained backbone network, enhancing the model's ability to capture shared information across modalities. Second, we incorporate a regional density supervision mechanism during the fine-tuning stage, which differentiates features in regions with varying crowd densities. Extensive experiments on three multi-modal crowd counting datasets validate our approach, making it the first to achieve an MAE below 10 on RGBT-CC. The code is avai

lable at https://github.com/HenryCilence/Free-Lunch-Multimodal-Counting.
****
BIMBA: Selective-Scan Compression for Long-Range Video Question Answering
Md Mohaiminul Islam, Tushar Nagarajan, Huiyu Wang, Gedas Bertasius, Lorenzo Torresani; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29096-29107
Video Question Answering (VQA) in long videos poses the key challenge of extracting relevant information and modeling long-range dependencies from many redundant frames. The self-attention mechanism provides a general solution for sequence modeling, but it has a prohibitive cost when applied to a massive number of spatiotemporal tokens in long videos. Most prior methods rely on compression strategies to lower the computational cost, such as reducing the input length via sparse frame sampling or compressing the output sequence passed to the large language model (LLM) via space-time pooling. However, these naive approaches over-represent redundant information and often miss salient events or fast-occurring space-time patterns. In this work, we introduce BIMBA, an efficient state-space model to handle long-form videos. Our model leverages the selective scan algorithm to learn to effectively select critical information from high-dimensional video and transform it into a reduced token sequence for efficient LLM processing. Extensive experiments demonstrate that BIMBA achieves state-of-the-art accuracy on multiple long-form VQA benchmarks, including PerceptionTest, NExT-QA, EgoSchema, VNBench, LongVideoBench, Video-MME, and MLVU.
****
EVolSplat: Efficient Volume-based Gaussian Splatting for Urban View Synthesis
Sheng Miao, Jiaxin Huang, Dongfeng Bai, Xu Yan, Hongyu Zhou, Yue Wang, Bingbing Liu, Andreas Geiger, Yiyi Liao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11286-11296
Novel view synthesis of urban scenes is essential for autonomous driving-related applications. Existing NeRF and 3DGS-based methods show promising results in achieving photorealistic renderings but require slow, per-scene optimization. We introduce EVolSplat, an efficient 3D Gaussian Splatting model for urban scenes that works in a feed-forward manner. Unlike existing feed-forward, pixel-aligned 3DGS methods, which often suffer from issues like multi-view inconsistencies and duplicated content, our approach predicts 3D Gaussians across multiple frames within a unified volume using a 3D convolutional network. This is achieved by initializing 3D Gaussians with noisy depth predictions, and then refining their geometric properties in 3D space and predicting color based on 2D textures. Our model also handles distant views and the sky with a flexible hemisphere background model. This enables us to perform fast, feed-forward reconstruction while achieving real-time rendering. Experimental evaluations on the KITTI-360 and Waymo datasets show that our method achieves state-of-the-art quality compared to existing feed-forward 3DGS- and NeRF-based methods.
****
Diff2Flow: Training Flow Matching Models via Diffusion Model Alignment
Johannes Schusterbauer, Ming Gui, Frank Fundel, Björn Ommer; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28347-28357
Diffusion models have revolutionized generative tasks through high-fidelity outputs, yet flow matching (FM) offers faster inference and empirical performance gains. However, current foundation FM models are computationally prohibitive for finetuning, while diffusion models like Stable Diffusion benefit from efficient architectures and ecosystem support. This work addresses the critical challenge of efficiently transferring knowledge from pre-trained diffusion models to flow matching. We propose Diff2Flow, a novel framework that systematically bridges diffusion and FM paradigms by rescaling timesteps, aligning interpolants, and deriving FM-compatible velocity fields from diffusion predictions. This alignment enables direct and efficient FM finetuning of diffusion priors with no extra computation overhead. Our experiments demonstrate that Diff2Flow outperforms naive FM and diffusion finetuning particularly under parameter-efficient constraints, while achieving superior or competitive performance across diverse downstream tasks compared to state-of-the-art methods.

****
JanusFlow: Harmonizing Autoregression and Rectified Flow for Unified Multimodal Understanding and Generation
Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, Liang Zhao, Yisong Wang, Jiaying Liu, Chong Ruan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7739-7751
We present JanusFlow, a powerful framework that unifies image understanding and generation in a single model.JanusFlow introduces a minimalist architecture that integrates autoregressive language models with rectified flow, a state-of-the-art method in generative modeling.Our key finding demonstrates that rectified flow can be straightforwardly trained within the large language model framework, eliminating the need for complex architectural modifications.To further improve the performance of our unified model, we adopt two key strategies: (i) decoupling the understanding and generation encoders, and (ii) aligning their representations during unified training.Extensive experiments show that JaunsFlow achieves comparable or superior performance to specialized models in their respective domains, while significantly outperforming existing unified approaches.This work represents a step toward more efficient and versatile vision-language models.
****
Visual Prompting for One-shot Controllable Video Editing without Inversion
Zhengbo Zhang, Yuxi Zhou, Duo Peng, Joo-Hwee Lim, Zhigang Tu, De Wen Soh, Lin Geng Foo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7784-7794
One-shot controllable video editing (OCVE) is an important yet challenging task, aiming to propagate user edits that are made---using any image editing tool---on the first frame of a video to all subsequent frames, while ensuring content consistency between edited frames and source frames. To achieve this, prior methods employ DDIM inversion to transform source frames into latent noise, which is then fed into a pre-trained diffusion model, conditioned on the user-edited first frame, to generate the edited video. However, the DDIM inversion process accumulates errors, which hinder the latent noise from accurately reconstructing the source frames, ultimately compromising content consistency in the generated edited frames. To overcome it, our method eliminates the need for DDIM inversion by performing OCVE through a novel perspective based on visual prompting. Furthermore, inspired by consistency models that can perform multi-step consistency sampling to generate a sequence of content-consistent images, we propose a content consistency sampling (CCS) to ensure content consistency between the generated edited frames and the source frames. Moreover, we introduce a temporal-content consistency sampling (TCS) based on Stein Variational Gradient Descent to ensure temporal consistency across the edited frames. Extensive experiments validate the effectiveness of our approach.
****
PIDSR: Complementary Polarized Image Demosaicing and Super-Resolution
Shuangfan Zhou, Chu Zhou, Youwei Lyu, Heng Guo, Zhanyu Ma, Boxin Shi, Imari Sato; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16081-16090
Polarization cameras can capture multiple polarized images with different polarizer angles in a single shot, bringing convenience to polarization-based downstream tasks. However, their direct outputs are color-polarization filter array (CPFA) raw images, requiring demosaicing to reconstruct full-resolution, full-color polarized images; unfortunately, this necessary step introduces artifacts that make polarization-related parameters such as the degree of polarization (DoP) and angle of polarization (AoP) prone to error. Besides, limited by the hardware design, the resolution of a polarization camera is often much lower than that of a conventional RGB camera. Existing polarized image demosaicing (PID) methods are limited in that they cannot enhance resolution, while polarized image super-resolution (PISR) methods, though designed to obtain high-resolution (HR) polarized images from the demosaicing results, tend to retain or even amplify errors in the DoP and AoP introduced by demosaicing artifacts. In this paper, we propose PI

DSR, a joint framework that performs complementary Polarized Image Demosaicing a
nd Super-Resolution, showing the ability to robustly obtain high-quality HR pola
rized images with more accurate DoP and AoP from a CPFA raw image in a direct ma
nner. Experiments show our PIDSR not only achieves state-of-the-art performance
on both synthetic and real data, but also facilitates downstream tasks.
****

MegaSynth: Scaling Up 3D Scene Reconstruction with Synthesized Data
Hanwen Jiang, Zexiang Xu, Desai Xie, Ziwen Chen, Haian Jin, Fujun Luan, Zhixin S
hu, Kai Zhang, Sai Bi, Xin Sun, Jiuxiang Gu, Qixing Huang, Georgios Pavlakos, Ha
o Tan; Proceedings of the Computer Vision and Pattern Recognition Conference (CV
PR), 2025, pp. 16441-16452
We propose scaling up 3D scene reconstruction by training with synthesized data.
 At the core of our work is MegaSynth, a procedurally generated 3D dataset compr
ising 700K scenes - over 50 times larger than the prior real dataset DL3DV - dra
matically scaling the training data. To enable scalable data generation, our key
 idea is eliminating semantic information, removing the need to model complex se
mantic priors such as object affordances and scene composition. Instead, we mode
l scenes with basic spatial structures and geometry primitives, offering scalabi
lity. Besides, we control data complexity to facilitate training while loosely a
ligning it with real-world data distribution to benefit real-world generalizatio
n. We explore training LRMs with both MegaSynth and available real data. Experim
ent results show that joint training or pre-training with MegaSynth improves rec
onstruction quality by 1.2 to 1.8 dB PSNR across diverse image domains. Moreover
, models trained solely on MegaSynth perform comparably to those trained on real
 data, underscoring the low-level nature of 3D reconstruction. Additionally, we
provide an in-depth analysis of MegaSynth's properties for enhancing model capab
ility, training stability, and generalization.
****

Prof. Robot: Differentiable Robot Rendering Without Static and Self-Collisions
Quanyuan Ruan, Jiabao Lei, Wenhao Yuan, Yanglin Zhang, Dekun Lu, Guiliang Liu, K
ui Jia; Proceedings of the Computer Vision and Pattern Recognition Conference (C
VPR), 2025, pp. 22562-22572
Differentiable rendering has gained significant attention in the field of roboti
cs, with differentiable robot rendering emerging as an effective paradigm for le
arning robotic actions from image-space supervision. However, the lack of physic
al world perception in this approach may lead to potential collisions during act
ion optimization. In this work, we introduce a novel improvement on previous eff
orts by incorporating physical awareness of collisions through the learning of a
 neural robotic collision classifier. This enables the optimization of actions t
hat avoid collisions with static, non-interactable environments as well as the r
obot itself. To facilitate effective gradient optimization with the classifier,
we identify the underlying issue and propose leveraging Eikonal regularization t
o ensure consistent gradients for optimization. Our solution can be seamlessly i
ntegrated into existing differentiable robot rendering frameworks, utilizing gra
dients for optimization and providing a foundation for future applications of di
fferentiable rendering in robotics with improved reliability of interactions wit
h the physical world. Both qualitative and quantitative experiments demonstrate
the necessity and effectiveness of our method compared to previous solutions.
****

AVQACL: A Novel Benchmark for Audio-Visual Question Answering Continual Learning
Kaixuan Wu, Xinde Li, Xinling Li, Chuanfei Hu, Guoliang Wu; Proceedings of the C
omputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3252-3261
In this paper, a novel benchmark for audio-visual question answering continual l
earning (AVQACL) is introduced, aiming to study fine-grained scene understanding
 and spatial-temporal reasoning in videos under a continual learning setting. To
 facilitate this multimodal continual leaning task, we create two audio-visual q
uestion answering continual learning datasets, named Split-AVQA and Split-MUSIC-
AVQA based on the AVQA and MUSIC-AVQA datasets, respectively. The experimental r
esults suggest that the model exhibits limited cognitive and reasoning abilities
 and experiences catastrophic forgetting when processing three modalities simult

aneously in a continuous data stream. To address above challenges, we propose a novel continual learning method that incorporates question-guided cross-modal information fusion (QCIF) to focus on question-relevant details for improved feature representation and task-specific knowledge distillation with spatial-temporal feature constraints (TKD-STFC) to preserve the spatial-temporal reasoning knowledge acquired from previous dynamic scenarios. Furthermore, a question semantic consistency constraint (QSCC) is employed to ensure that the model maintains a consistent understanding of question semantics across tasks throughout the continual learning process. Extensive experimental results on Split-AVQA and Split-MUSIC-AVQA datasets illustrate that our method achieves state-of-the-art audio-visual question answering continual learning performance. The code is available at https://github.com/kx-wu/CVPR2025_AVQACL.

****

Flash-Split: 2D Reflection Removal with Flash Cues and Latent Diffusion Separation

Tianfu Wang, Mingyang Xie, Haoming Cai, Sachin Shah, Christopher A. Metzler; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5688-5698

Transparent surfaces, such as glass, create complex reflections that obscure images and challenge downstream computer vision applications. We introduce Flash-Split, a robust framework for separating transmitted and reflected light using a single (potentially misaligned) pair of flash/no-flash images. Our core idea is to perform flash-informed reflection separation iteratively in a low-dimensional latent space. Specifically, Flash-Split consists of two stages. Stage 1 separates the reflection latent and transmission latent via a dual-branch diffusion model that is conditioned on an encoded flash/no-flash latent pair. This stage effectively mitigates the flash/no-flash misalignment issue. Stage 2 restores high-resolution, faithful details to the separated latents via a cross-latent decoding process that is conditioned on the original images before separation. We validate Flash-Split on challenging real-world scenes and demonstrate it significantly outperforms existing methods. Our project webpage is at https://flash-split.github.io/.

****

Attention IoU: Examining Biases in CelebA using Attention Maps

Aaron Serianni, Tyler Zhu, Olga Russakovsky, Vikram V. Ramaswamy; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4386-4397

Computer vision models have been shown to exhibit and amplify biases across a wide array of datasets and tasks. Existing methods for quantifying bias in classification models primarily focus on dataset distribution and model performance on subgroups, overlooking the internal workings of a model. We introduce the Attention-IoU (Attention Intersection over Union) metric and related scores, which use attention maps to reveal biases within a model's internal representations and identify image features potentially causing the biases. First, we validate Attention-IoU on the synthetic Waterbirds dataset, showing that the metric accurately measures model bias. We then analyze the CelebA dataset, finding that Attention-IoU uncovers correlations beyond accuracy disparities. Through an investigation of individual attributes through the protected attribute of Male, we examine the distinct ways biases are represented in CelebA. Lastly, by subsampling the training set to change attribute correlations, we demonstrate that Attention-IoU reveals potential confounding variables not present in dataset labels. Our code is available at https://github.com/aaronserianni/attention-iou.

****

HEIE: MLLM-Based Hierarchical Explainable AIGC Image Implausibility Evaluator

Fan Yang, Ru Zhen, Jianing Wang, Yanhao Zhang, Haoxiang Chen, Haonan Lu, Sicheng Zhao, Guiguang Ding; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3856-3866

AIGC images are prevalent across various fields, yet they frequently suffer from quality issues like artifacts and unnatural textures. Specialized models aim to predict defect region heatmaps but face two primary challenges: (1) lack of exp

lainability, failing to provide reasons and analyses for subtle defects, and (2) inability to leverage common sense and logical reasoning, leading to poor generalization. Multimodal large language models (MLLMs) promise better comprehension and reasoning but face their own challenges: (1) difficulty in fine-grained defect localization due to the limitations in capturing tiny details, and (2) constraints in providing pixel-wise outputs necessary for precise heatmap generation. To address these challenges, we propose HEIE: a novel MLLM-Based Hierarchical Explainable Image Implausibility Evaluator. We introduce the CoT-Driven Explainable Trinity Evaluator, which integrates heatmaps, scores, and explanation outputs, using CoT to decompose complex tasks into subtasks of increasing difficulty and enhance interpretability. Our Adaptive Hierarchical Implausibility Mapper synergizes low-level image features with high-level mapper tokens from LLMs, enabling precise local-to-global hierarchical heatmap predictions through an uncertainty-based adaptive token approach. Moreover, we propose a new dataset: Expl-AIGI-Eval, designed to facilitate interpretable implausibility evaluation of AIGC images. Our method demonstrates state-of-the-art performance through extensive experiments. Our project is at https://yfthu.github.io/HEIE/.
****

Segment Any Motion in Videos
Nan Huang, Wenzhao Zheng, Chenfeng Xu, Kurt Keutzer, Shanghang Zhang, Angjoo Kanazawa, Qianqian Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3406-3416
Moving object segmentation is a crucial task for achieving a high-level understanding of visual scenes and has numerous downstream applications. Humans can effortlessly segment moving objects in videos. Previous work has largely relied on optical flow to provide motion cues; however, this approach often results in imperfect predictions due to challenges such as partial motion, complex deformations, motion blur and background distractions. We propose a novel approach for moving object segmentation that combines long-range trajectory motion cues with DINO-based semantic features and leverages SAM2 for pixel-level mask densification through an iterative prompting strategy. Our model employs Spatio-Temporal Trajectory Attention and Motion-Semantic Decoupled Embedding to prioritize motion while integrating semantic support. Extensive testing on diverse datasets demonstrates state-of-the-art performance, excelling in challenging scenarios and fine-grained segmentation of multiple objects. Our code is available at https://motion-seg.github.io/.
****

HandOS: 3D Hand Reconstruction in One Stage
Xingyu Chen, Zhuheng Song, Xiaoke Jiang, Yaoqing Hu, Junzhi Yu, Lei Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17304-17314
Existing approaches of hand reconstruction predominantly adhere to a multi-stage framework, encompassing detection, left-right classification, and pose estimation. This paradigm induces redundant computation and cumulative errors. In this work, we propose HandOS, an end-to-end framework for 3D hand reconstruction. Our central motivation lies in leveraging a frozen detector as the foundation while incorporating auxiliary modules for 2D and 3D keypoint estimation. In this manner, we integrate the pose estimation capacity into the detection framework, while at the same time obviating the necessity of using the left-right category as a prerequisite. Specifically, we propose an interactive 2D-3D decoder, where 2D joint semantics is derived from detection cues while 3D representation is lifted from those of 2D joints. Furthermore, hierarchical attention is designed to enable the concurrent modeling of 2D joints, 3D vertices, and camera translation. Consequently, we achieve an end-to-end integration of hand detection, 2D pose estimation, and 3D mesh reconstruction within a one-stage framework, so that the above multi-stage drawbacks are overcome. Meanwhile, the HandOS reaches state-of-the-art performances on public benchmarks, e.g., 5.0 PA-MPJPE on FreiHand and 64.6% PCK@0.05 on HInt-Ego4D.
****

Task-aware Cross-modal Feature Refinement Transformer with Large Language Models

for Visual Grounding
Wenbo Chen, Zhen Xu, Ruotao Xu, Si Wu, Hau-San Wong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3931-3941
The goal of visual grounding is to establish connections between target objects and textual descriptions. Large Language Models (LLMs) have demonstrated strong comprehension abilities across a variety of visual tasks. To establish precise associations between the text and the corresponding visual region, we propose a Task-aware Cross-modal feature Refinement Transformer with LLMs for visual grounding, and our model is referred to as TCRT. To enable the LLM trained solely on text to understand images, we introduce an LLM adaptation module that extracts text-related visual features to bridge the domain discrepancy between the textual and visual modalities. We feed the text and visual features into the LLM to obtain task-aware priors. To enable the priors to guide feature fusion process, we further incorporate a cross-modal feature fusion module, which allows task-aware embeddings to refine visual features and facilitate information interaction between the Referring Expression Comprehension (REC) and Referring Expression Segmentation (RES) tasks. We have performed extensive experiments to verify the effectiveness of the main components and demonstrate the superior performance of the proposed TCRT over state-of-the-art end-to-end visual grounding methods on RefCOCO, RefCOCOg, RefCOCO+ and ReferItGame.
****
DropGaussian: Structural Regularization for Sparse-view Gaussian Splatting
Hyunwoo Park, Gun Ryu, Wonjun Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21600-21609
Recently, 3D Gaussian splatting (3DGS) has gained considerable attentions in the field of novel view synthesis due to its fast performance while yielding the excellent image quality. However, 3DGS in sparse-view settings (e.g., three-view inputs) often faces with the problem of overfitting to training views, which significantly drops the visual quality of novel view images. Many existing approaches have tackled this issue by using strong priors, such as 2D generative contextual information and external depth signals. In contrast, this paper introduces a prior-free method, so-called DropGaussian, with simple changes in 3D Gaussian splatting. Specifically, we randomly remove Gaussians during the training process in a similar way of dropout, which allows non-excluded Gaussians to have larger gradients while improving their visibility. This makes the remaining Gaussians to contribute more to the optimization process for rendering with sparse input views. Such simple operation effectively alleviates the overfitting problem and enhances the quality of novel view synthesis. By simply applying DropGaussian to the original 3DGS framework, we can achieve the competitive performance with existing prior-based 3DGS methods in sparse-view settings of benchmark datasets without any additional complexity.
****
All-Day Multi-Camera Multi-Target Tracking
Huijie Fan, Yu Qiao, Yihao Zhen, Tinghui Zhao, Baojie Fan, Qiang Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16892-16901
The capability of tracking objects in low-light environments like nighttime is crucial for numerous real-world applications such as crowd behavior analysis and traffic scene understanding. However, previous Multi-Camera Multi-Target(MCMT) tracking methods are primarily focused on tracking during daytime with favorable lighting, shying away from low-light environments. The main difficulty of tracking under low light condition is lack of detailed visible appearance features. To address this issue, we incorporate the infrared modality into MCMT tracking framework to provide more useful information. We constructed the first Multi-modality(RGBT) Multi-camera Multi-target tracking dataset named M3Track, which contains sequences captured in low-light environments, laying a solid foundation for all-day multi-camera tracking. Based on the proposed dataset, we propose All-Day Multi-Camera Multi-Target tracking network, termed as ADMCMT. Specifically, we propose an All-Day Mamba Fusion(ADMF) model to adaptively fuse information from different modalities. Within ADMF, the Lighting Guidance Model(IGM) extracts light

ing relevant information to guide the fusion process. Furthermore, the Nearby Ta
rget Collection(NTC) strategy is designed to enhance tracking accuracy by levera
ging information derived from surrounding objects of target. Experiments conduct
ed on M3Track demonstrate that ADMCMT exhibits strong generalization across diff
erent lighting conditions. The code will be released soon at https://github.com/
QTRACKY/ADMCMT.
****

Blurred LiDAR for Sharper 3D: Robust Handheld 3D Scanning with Diffuse LiDAR and
 RGB

Nikhil Behari, Aaron Young, Siddharth Somasundaram, Tzofi Klinghoffer, Akshat Da
ve, Ramesh Raskar; Proceedings of the Computer Vision and Pattern Recognition Co
nference (CVPR), 2025, pp. 26954-26964

3D surface reconstruction is essential across applications of virtual reality, r
obotics, and mobile scanning. However, RGB-based reconstruction often fails in l
ow-texture, low-light, and low-albedo scenes. Handheld LiDARs, now common on mob
ile devices, aim to address these challenges by capturing depth information from
 time-of-flight measurements of a coarse grid of projected dots. Yet, these spar
se LiDARs struggle with scene coverage on limited input views, leaving large gap
s in depth information. In this work, we propose using an alternative class of "
blurred" LiDAR that emits a diffuse flash, greatly improving scene coverage but
introducing spatial ambiguity from mixed time-of-flight measurements across a wi
de field of view. To handle these ambiguities, we propose leveraging the complem
entary strengths of diffuse LiDAR with RGB. We introduce a Gaussian surfel-based
 rendering framework with a scene-adaptive loss function that dynamically balanc
es RGB and diffuse LiDAR signals. We demonstrate that, surprisingly, diffuse LiD
AR can outperform traditional sparse LiDAR, enabling robust 3D scanning with acc
urate color and geometry estimation in challenging environments.
****

EnergyMoGen: Compositional Human Motion Generation with Energy-Based Diffusion M
odel in Latent Space

Jianrong Zhang, Hehe Fan, Yi Yang; Proceedings of the Computer Vision and Patter
n Recognition Conference (CVPR), 2025, pp. 17592-17602

Diffusion models, particularly latent diffusion models, have demonstrated remark
able success in text-driven human motion generation. However, it remains challen
ging for latent diffusion models to effectively compose multiple semantic concep
ts into a single, coherent motion sequence. To address this issue, we propose En
ergyMoGen, which includes two spectrums of Energy-Based Models: (1) We interpret
 the diffusion model as a latent-aware energy-based model that generates motions
 by composing a set of diffusion models in latent space; (2) We introduce a sema
ntic-aware energy model based on cross-attention, which enables semantic composi
tion and adaptive gradient descent for text embeddings. To overcome the challeng
es of semantic inconsistency and motion distortion across these two spectrums, w
e introduce Synergistic Energy Fusion. This design allows the motion latent diff
usion model to synthesize high-quality, complex motions by combining multiple en
ergy terms corresponding to textual descriptions. Experiments show that our appr
oach outperforms existing state-of-the-art models on various motion generation t
asks, including text-to-motion generation, compositional motion generation, and
multi-concept motion generation. Additionally, we demonstrate that our method ca
n be used to extend motion datasets and improve the text-to-motion task.
****

PatchDEMUX: A Certifiably Robust Framework for Multi-label Classifiers Against A
dversarial Patches

Dennis Jacob, Chong Xiang, Prateek Mittal; Proceedings of the Computer Vision an
d Pattern Recognition Conference (CVPR), 2025, pp. 9944-9953

Deep learning techniques have enabled vast improvements in computer vision techn
ologies. Nevertheless, these models are vulnerable to adversarial patch attacks
which catastrophically impair performance. The physically realizable nature of
these attacks calls for certifiable defenses, which feature provable guarantees
on robustness. While certifiable defenses have been successfully applied to sing
le-label classification, limited work has been done for multi-label classificati

on. In this work, we present PatchDEMUX, a certifiably robust framework for multi-label classifiers against adversarial patches. Our approach is a generalizable method which can extend any existing certifiable defense for single-label classification; this is done by considering the multi-label classification task as a series of isolated binary classification problems to provably guarantee robustness. Furthermore, in the scenario where an attacker is limited to a single patch we propose an additional certification procedure that can provide tighter robustness bounds. Using the current state-of-the-art (SOTA) single-label certifiable defense PatchCleanser as a backbone, we find that PatchDEMUX can achieve non-trivial robustness on the MS-COCO and PASCAL VOC datasets while maintaining high clean performance

****

StarVector: Generating Scalable Vector Graphics Code from Images and Text

Juan A. Rodriguez, Abhay Puri, Shubham Agarwal, Issam H. Laradji, Pau Rodriguez, Sai Rajeswar, David Vazquez, Christopher Pal, Marco Pedersoli; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16175-16186

Scalable Vector Graphics (SVGs) are vital for modern image rendering due to their scalability and versatility. Previous SVG generation methods have focused on curve-based vectorization, lacking semantic understanding, often producing artifacts, and struggling with SVG primitives beyond path curves. To address these issues, we introduce StarVector, a multimodal large language model for SVG generation. It performs image vectorization by understanding image semantics and using SVG primitives for compact, precise outputs. Unlike traditional methods, StarVector works directly in the SVG code space, leveraging visual understanding to apply accurate SVG primitives. To train StarVector, we create SVG-Stack, a diverse dataset of 2M samples that enables generalization across vectorization tasks and precise use of primitives like ellipses, polygons, and text. We address challenges in SVG evaluation, showing that pixel-based metrics like MSE fail to capture the unique qualities of vector graphics. We introduce SVG-Bench, a benchmark across 10 datasets, and three tasks: image vectorization, text-driven SVG generation, and diagram generation. Using this contribution, StarVector achieves state-of-the-art performance, producing more compact and semantically rich SVGs.

****

Novel View Synthesis with Pixel-Space Diffusion Models

Noam Elata, Bahjat Kawar, Yaron Ostrovsky-Berman, Miriam Farber, Ron Sokolovsky; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26756-26766

Synthesizing a novel view from a single input image is a challenging task. Traditionally, this task was approached by estimating scene depth, warping, and inpainting, with machine learning models enabling parts of the pipeline. More recently, generative models are being increasingly employed in novel view synthesis (NVS), often encompassing the entire end-to-end system.In this work, we adapt a modern diffusion model architecture for end-to-end NVS in the pixel space, substantially outperforming previous state-of-the-art (SOTA) techniques. We explore different ways to encode geometric information into the network.Our experiments show that while these methods may enhance performance, their impact is minor compared to utilizing improved generative models. Moreover, we introduce a novel NVS training scheme that utilizes single-view datasets, capitalizing on their relative abundance compared to their multi-view counterparts. This leads to improved generalization capabilities to scenes with out-of-domain content.

****

Object Detection using Event Camera: A MoE Heat Conduction based Detector and A New Benchmark Dataset

Xiao Wang, Yu Jin, Wentao Wu, Wei Zhang, Lin Zhu, Bo Jiang, Yonghong Tian; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29321-29330

Object detection in event streams has emerged as a cutting-edge research area, demonstrating superior performance in low-light conditions, scenarios with motion blur, and rapid movements. Current detectors leverage spiking neural networks,

Transformers, or convolutional neural networks as their core architectures, each with its own set of limitations including restricted performance, high computational overhead, or limited local receptive fields. This paper introduces a novel MoE (Mixture of Experts) heat conduction-based object detection algorithm that strikingly balances accuracy and computational efficiency. Initially, we employ a stem network for event data embedding, followed by processing through our innovative MoE-HCO blocks. Each block integrates various expert modules to mimic heat conduction within event streams. Subsequently, an IoU-based query selection module is utilized for efficient token extraction, which is then channeled into a detection head for the final object detection process. Furthermore, we are pleased to introduce EvDET200K, a novel benchmark dataset for event-based object detection. Captured with a high-definition Prophesee EVK4-HD event camera, this dataset encompasses 10 distinct categories, 200,000 bounding boxes, and 10,054 samples, each spanning 2 to 5 seconds. We also provide comprehensive results from over 15 state-of-the-art detectors, offering a solid foundation for future research and comparison.
****

EgoTextVQA: Towards Egocentric Scene-Text Aware Video Question Answering
Sheng Zhou, Junbin Xiao, Qingyun Li, Yicong Li, Xun Yang, Dan Guo, Meng Wang, Tat-Seng Chua, Angela Yao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3363-3373
We introduce EgoTextVQA, a novel and rigorously constructed benchmark for egocentric QA assistance involving scene text. EgoTextVQA contains 1.5K ego-view videos and 7K scene-text aware questions that reflect real user needs in outdoor driving and indoor house-keeping activities. The questions are designed to elicit identification and reasoning on scene text in an egocentric and dynamic environment. With EgoTextVQA, we comprehensively evaluate 10 prominent multimodal large language models. Currently, all models struggle, and the best results (Gemini 1.5 Pro) are around 33% accuracy, highlighting the severe deficiency of these techniques in egocentric QA assistance. Our further investigations suggest that precise temporal grounding and multi-frame reasoning, along with high resolution and auxiliary scene-text inputs, are key for better performance. With thorough analyses and heuristic suggestions, we hope EgoTextVQA can serve as a solid testbed for research in egocentric scene-text QA assistance.
****

Token Cropr: Faster ViTs for Quite a Few Tasks
Benjamin Bergner, Christoph Lippert, Aravindh Mahendran; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9740-9750
The adoption of Vision Transformers (ViTs) in resource-constrained applications necessitates improvements in inference throughput. To this end several token pruning and merging approaches have been proposed that improve efficiency by successively reducing the number of tokens. However, it remains an open problem to design a token reduction method that is fast, maintains high performance, and is applicable to various vision tasks. In this work, we present a token pruner that uses auxiliary prediction heads that learn to select tokens end-to-end based on task relevance. These auxiliary heads can be removed after training, leading to throughput close to that of a random pruner. We evaluate our method on image classification, semantic segmentation, object detection, and instance segmentation, and show speedups of 1.5 to 4x with small drops in performance. As a best case, on the ADE20k semantic segmentation benchmark, we observe a 2x speedup relative to the no-pruning baseline, with a negligible performance penalty of 0.1 median mIoU across 5 seeds.
****

STCOcc: Sparse Spatial-Temporal Cascade Renovation for 3D Occupancy and Scene Flow Prediction
Zhimin Liao, Ping Wei, Shuaijia Chen, Haoxuan Wang, Ziyang Ren; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1516-1526
3D occupancy and scene flow offer a detailed and dynamic representation of 3D scene. Recognizing the sparsity and complexity of 3D space, previous vision-centri

c methods have employed implicit learning-based approaches to model spatial and temporal information. However, these approaches struggle to capture local details and diminish the model's spatial discriminative ability. To address these challenges, we propose a novel explicit state-based modeling method designed to leverage the occupied state to renovate the 3D features. Specifically, we propose a sparse occlusion-aware attention mechanism, integrated with a cascade refinement strategy, which accurately renovates 3D features with the guidance of occupied state information. Additionally, we introduce a novel method for modeling long-term dynamic interactions, which reduces computational costs and preserves spatial information. Compared to the previous state-of-the-art methods, our efficient explicit renovation strategy not only delivers superior performance in terms of RayIoU and mAVE for occupancy and scene flow prediction but also markedly reduces GPU memory usage during training, bringing it down to 8.7GB. Our code is available on https://github.com/lzzzzzm/STCOcc

****

Document Haystacks: Vision-Language Reasoning Over Piles of 1000+ Documents

Jun Chen, Dannong Xu, Junjie Fei, Chun-Mei Feng, Mohamed Elhoseiny; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24817-24826

Large multimodal models (LMMs) have achieved impressive progress in vision-language understanding, yet they face limitations in real-world applications requiring complex reasoning over a large number of images. Existing benchmarks for multi-image question-answering are limited in scope, each question is paired with only up to 30 images, which does not fully capture the demands of large-scale retrieval tasks encountered in the real-world usages. To reduce these gaps, we introduce two document haystack benchmarks, dubbed DocHaystack and InfoHaystack, designed to evaluate LMM performance on large-scale visual document retrieval and understanding. Additionally, we propose V-RAG, a novel, vision-centric retrieval-augmented generation (RAG) framework that leverages a suite of multimodal vision encoders, each optimized for specific strengths, and a dedicated question-document relevance module. V-RAG sets a new standard, with a 9% and 11% improvement in Recall@1 on the challenging DocHaystack-1000 and InfoHaystack-1000 benchmarks, respectively, compared to the previous best baseline models. Additionally, integrating V-RAG with LMMs enables them to efficiently operate across thousands of images, yielding significant improvements on our DocHaystack and InfoHaystack benchmarks. Our code and datasets will be made publicly available.

****

Rethinking Few-Shot Adaptation of Vision-Language Models in Two Stages

Matteo Farina, Massimiliano Mancini, Giovanni Iacca, Elisa Ricci; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29989-29998

An old-school recipe for training a classifier is to (i) learn a good feature extractor and (ii) optimize a linear layer atop. When only a handful of samples are available per category, as in Few-Shot Adaptation (FSA), data are insufficient to fit a large number of parameters, rendering the above impractical. This is especially true with large pre-trained Vision-Language Models (VLMs), which motivated successful research at the intersection of Parameter-Efficient Fine-tuning (PEFT) and FSA. In this work, we start by analyzing the learning dynamics of PEFT techniques when trained on few-shot data from only a subset of categories, referred to as the "base" classes. We show that such dynamics naturally splits into two distinct phases: (i) task-level feature extraction and (ii) specialization to the available concepts. To accommodate this dynamic, we then depart from prompt- or adapter-based methods and tackle FSA differently. Specifically, given a fixed computational budget, we split it to (i) learn a task-specific feature extractor via PEFT and (ii) train a linear classifier on top. We call this scheme Two-Stage Few-Shot Adaptation (2SFS). Differently from established methods, our scheme enables a novel form of selective inference at a category level, i.e., at test time, only novel categories are embedded by the adapted text encoder, while embeddings of base categories are available within the classifier. Results with fixed hyperparameters across two settings, three backbones, and eleven datasets,

show that 2SFS matches or surpasses the state-of-the-art, while established methods degrade significantly across settings.
****

Resilient Sensor Fusion Under Adverse Sensor Failures via Multi-Modal Expert Fusion

Konyul Park, Yecheol Kim, Daehun Kim, Jun Won Choi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6720-6729

Modern autonomous driving perception systems utilize complementary multi-modal sensors, such as LiDAR and cameras. Although sensor fusion architectures enhance performance in challenging environments, they still suffer significant performance drops under severe sensor failures, such as LiDAR beam reduction, LiDAR drop, limited field of view, camera drop, and occlusion. This limitation stems from inter-modality dependencies in current sensor fusion frameworks. In this study, we introduce an efficient and robust LiDAR-camera 3D object detector, referred to as MoME, which can achieve robust performance through a mixture of experts approach. Our MoME fully decouples modality dependencies using three parallel expert decoders, which use camera features, LiDAR features, or a combination of both to decode object queries, respectively. We propose Multi-Expert Decoding (MED) framework, where each query is decoded selectively using one of three expert decoders. MoME utilizes an Adaptive Query Router (AQR) to select the most appropriate expert decoder for each query based on the quality of camera and LiDAR features. This ensures that each query is processed by the best-suited expert, resulting in robust performance across diverse sensor failure scenarios. We evaluated the performance of MoME on the nuScenes-R benchmark. Our MoME achieved state-of-the-art performance in extreme weather and sensor failure conditions, significantly outperforming the existing models across various sensor failure scenarios.
****

TAGA: Self-supervised Learning for Template-free Animatable Gaussian Articulated Model

Zhichao Zhai, Guikun Chen, Wenguan Wang, Dong Zheng, Jun Xiao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21159-21169

Decoupling from customized parametric templates represents a crucial step toward the creation of fully flexible, animatable articulated models. While existing template-free methods can achieve high-fidelity reconstruction in observed views, they struggle to recover plausible canonical models, resulting in suboptimal animation quality. This limitation stems from overlooking the fundamental ambiguity in canonical reconstruction, where multiple canonical models could explain the same observed views. By revealing the entanglement between the canonical ambiguity and incorrect skinning, we present a self-supervised framework that learns both plausible skinning and accurate canonical geometry using only sparse pose data. Our method, TAGA, uses explicit 3D Gaussians as skinning carriers and characterizes the ambiguity as "Ambiguous Gaussians" with incorrect skinning weights. TAGA then corrects ambiguous Gaussians in the observation space using anomaly detection. With the corrected ones, we enforce cycle consistency constraints on both geometry and skinning to refine the corresponding Gaussians in the canonical space through a new backward method. Compared to existing state-of-the-art template-free methods, TAGA delivers superior visual fidelity for novel views and poses, while significantly improving training and rendering speeds. Experiments on challenging datasets with limited pose variations further demonstrate the robustness and generality of TAGA.
****

MambaVO: Deep Visual Odometry Based on Sequential Matching Refinement and Training Smoothing

Shuo Wang, Wanting Li, Yongcai Wang, Zhaoxin Fan, Zhe Huang, Xudong Cai, Jian Zhao, Deying Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1252-1262

Deep visual odometry has demonstrated great advancements by learning-to-optimize technology. This approach heavily relies on the visual matching across frames. However, ambiguous matching in challenging scenarios leads to significant error

s in geometric modeling and bundle adjustment optimization, which undermines the accuracy and robustness of pose estimation. To address this challenge, this paper proposes MambaVO, which conducts robust initialization, Mamba-based sequential matching refinement, and smoothed training to enhance the matching quality and improve the pose estimation in deep visual odometry. Specifically, when a new frame is received, it is matched with the closest keyframe in the maintained Point-Frame Graph (PFG) via the semi-dense based Geometric Initialization Module (GIM). Then the initialized PFG is processed by a proposed Geometric Mamba Module (GMM), which exploits the matching features to refine the overall inter-frame pixel-to-pixel matching. The refined PFG is finally processed by deep BA to optimize the poses and the map. To deal with the gradient variance, a Trending-Aware Penalty (TAP) is proposed to smooth training by balancing the pose loss and the matching loss to enhance convergence and stability. A loop closure module is finally applied to enable MambaVO++. On public benchmarks, MambaVO and MambaVO++ demonstrate SOTA accuracy performance, while ensuring real-time running performance with low GPU memory requirement. Codes will be publicly available.

****

Explaining in Diffusion: Explaining a Classifier with Diffusion Semantics

Tahira Kazimi, Ritika Allada, Pinar Yanardag; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14799-14809

Classifiers are important components in many computer vision tasks, serving as the foundational backbone of a wide variety of models employed across diverse applications. However, understanding the decision-making process of classifiers remains a significant challenge. We propose DiffEx, a novel method that leverages the capabilities of text-to-image diffusion models to explain classifier decisions. Unlike traditional GAN-based explainability models, which are limited to simple, single-concept analyses and typically require training a new model for each classifier, our approach can explain classifiers that focus on single concepts (such as faces or animals) as well as those that handle complex scenes involving multiple concepts. DiffEx employs vision-language models to create a hierarchical list of semantics, allowing users to identify not only the overarching semantic influences on classifiers (e.g., the 'beard' semantic in a facial classifier) but also their sub-types, such as 'goatee' or 'Balbo' beard. Our experiments demonstrate that DiffEx is able to cover a significantly broader spectrum of semantics compared to its GAN counterparts, providing a hierarchical tool that delivers a more detailed and fine-grained understanding of classifier decisions.

****

Horizon-GS: Unified 3D Gaussian Splatting for Large-Scale Aerial-to-Ground Scenes

Lihan Jiang, Kerui Ren, Mulin Yu, Linning Xu, Junting Dong, Tao Lu, Feng Zhao, Dahua Lin, Bo Dai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26789-26799

Seamless integration of both aerial and street view images remains a significant challenge in neural scene reconstruction and rendering. Existing methods predominantly focus on single domain, limiting their applications in immersive environments, which demand extensive free view exploration with large view changes both horizontally and vertically. We introduce Horizon-GS, a novel approach built upon Gaussian Splatting techniques, tackles the unified reconstruction and rendering for aerial and street views. Our method addresses the key challenges of combining these perspectives with a new training strategy, overcoming viewpoint discrepancies to generate high-fidelity scenes. We also curated a high-quality aerial-to-ground view dataset encompassing both synthetic and real-world scene to advance further research. Experiments across diverse urban scene datasets confirms the effectiveness of our method.

****

Attention Distillation: A Unified Approach to Visual Characteristics Transfer

Yang Zhou, Xu Gao, Zichong Chen, Hui Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18270-18280

Recent advances in generative diffusion models have shown a notable inherent understanding of image style and semantics. In this paper, we leverage the self-att

ention features from pretrained diffusion networks to transfer the visual charac teristics from a reference to generated images. Unlike previous work that uses t hese features as plug-and-play attributes, we propose a novel attention distilla tion loss calculated between the ideal and current stylization results, based on which we optimize the synthesized image via backpropagation in latent space. Ne xt, we propose an improved Classifier Guidance that integrates attention distill ation loss into the denoising sampling process, further accelerating the synthes is and enabling a broad range of image generation applications. Extensive experi ments have demonstrated the extraordinary performance of our approach in transfe rring the examples' style, appearance, and texture to new images in synthesis.
****

From Words to Structured Visuals: A Benchmark and Framework for Text-to-Diagram Generation and Editing
Jingxuan Wei, Cheng Tan, Qi Chen, Gaowei Wu, Siyuan Li, Zhangyang Gao, Linzhuang Sun, Bihui Yu, Ruifeng Guo; Proceedings of the Computer Vision and Pattern Reco gnition Conference (CVPR), 2025, pp. 13315-13325
We introduce the task of text-to-diagram generation, which focuses on creating s tructured visual representations directly from textual descriptions. Existing ap proaches in text-to-image and text-to-code generation lack the logical organizat ion and flexibility needed to produce accurate, editable diagrams, often resulti ng in outputs that are either unstructured or difficult to modify. To address th is gap, we introduce DiagramGenBenchmark, a comprehensive evaluation framework e ncompassing eight distinct diagram categories, including flowcharts, model archi tecture diagrams, and mind maps. Additionally, we present DiagramAgent, an innov ative framework with four core modules--Plan Agent, Code Agent, Check Agent, and Diagram-to-Code Agent--designed to facilitate both the generation and refinemen t of complex diagrams. Our extensive experiments, which combine objective metric s with human evaluations, demonstrate that DiagramAgent significantly outperform s existing baseline models in terms of accuracy, structural coherence, and modif iability. This work not only establishes a foundational benchmark for the text-t o-diagram generation task but also introduces a powerful toolset to advance rese arch and applications in this emerging area.
****

LotusFilter: Fast Diverse Nearest Neighbor Search via a Learned Cutoff Table
Yusuke Matsui; Proceedings of the Computer Vision and Pattern Recognition Confer ence (CVPR), 2025, pp. 30430-30439
Approximate nearest neighbor search (ANNS) is an essential building block for ap plications like RAG but can sometimes yield results that are overly similar to e ach other. In certain scenarios, search results should be similar to the query a nd yet diverse. We propose LotusFilter, a post-processing module to diversify AN NS results. We precompute a cutoff table summarizing vectors that are close to e ach other. During the filtering, LotusFilter greedily looks up the table to dele te redundant vectors from the candidates. We demonstrated that the LotusFilter o perates fast (0.02 [ms/query]) in settings resembling real-world RAG application s, utilizing features such as OpenAI embeddings. Our code is publicly available at https://github.com/matsui528/lotf.
****

DreamRelation: Bridging Customization and Relation Generation
Qingyu Shi, Lu Qi, Jianzong Wu, Jinbin Bai, Jingbo Wang, Yunhai Tong, Xiangtai L i; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15723-15732
Customized image generation is essential for delivering personalized content bas ed on user-provided prompts, enabling large-scale text-to-image diffusion models to better align with individual needs. However, existing models often neglect t he relationships between customized objects in generated images. In contrast, th is work addresses this gap by focusing on relation-aware customized image genera tion, which seeks to preserve the identities from image prompts while maintainin g the predicate relations specified in text prompts. Specifically, we introduce DreamRelation, a framework that disentangles identity and relation learning usin g a carefully curated dataset. Our training data consists of relation-specific i

mages, independent object images containing identity information, and text prompts to guide relation generation. Then, we propose two key modules to tackle the two main challenges--generating accurate and natural relations, especially when significant pose adjustments are required, and avoiding object confusion in cases of overlap. First, we introduce a keypoint matching loss that effectively guides the model in adjusting object poses closely tied to their relationships. Second, we incorporate local features from the image prompts to better distinguish between objects, preventing confusion in overlapping cases. Extensive results on our proposed benchmarks demonstrate the superiority of DreamRelation in generating precise relations while preserving object identities across a diverse set of objects and relations. The source code and trained models will be made available to the public.

****

IndoorGS: Geometric Cues Guided Gaussian Splatting for Indoor Scene Reconstruction

Cong Ruan, Yuesong Wang, Tao Guan, Bin Zhang, Lili Ju; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 844-853

3D Gaussian Splatting (3DGS) has shown impressive performance in scene reconstruction, offering high rendering quality and rapid rendering speed with short training time. However, it often yields unsatisfactory results when applied to indoor scenes due to its poor ability to learn geometries without enough textural information. In this paper, we propose a new 3DGS-based method "IndoorGS", that leverages the commonly found yet important geometric cues in indoor scenes to improve the reconstruction quality. Specifically, we first extract 2D lines from input images and fuse them into 3D line cues via feature-based matching, which can provide a structural understanding of the target scene. We then apply the statistical outlier removal to refine Structure-from-Motion (SfM) points, ensuring robust cues in texture-rich areas. Based on these two types of cues, we further extract reliable 3D plane cues for textureless regions. Such geometric information will be utilized not only for initialization but also in the realization of a geometric-cue-guided adaptive density control (ADC) strategy. The proposed ADC approach is grounded in the principle of divide-and-conquer and optimizes the use of each type of geometric cues to enhance overall reconstruction performance. Extensive experiments on multiple indoor datasets show that our method can deliver much more accurate geometry and higher rendering quality for indoor scenes than existing 3DGS approaches.

****

Point-Cache: Test-time Dynamic and Hierarchical Cache for Robust and Generalizable Point Cloud Analysis

Hongyu Sun, Qiuhong Ke, Ming Cheng, Yongcai Wang, Deying Li, Chenhui Gou, Jianfei Cai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1263-1275

This paper proposes a general solution to enable point cloud recognition models to handle distribution shifts at test time. Unlike prior methods, which rely heavily on training data (often inaccessible during online inference) and are limited to recognizing a fixed set of point cloud classes predefined during training, we explore a more practical and challenging scenario: adapting the model solely based on online test data to recognize both previously seen classes and novel, unseen classes at test time. To this end, we develop Point-Cache, a hierarchical cache model that captures essential clues of online test samples, particularly focusing on the global structure of point clouds and their local-part details. Point-Cache, which serves as a rich 3D knowledge base, is dynamically managed to prioritize the inclusion of high-quality samples. Designed as a plug-and-play module, our method can be flexibly integrated into large multimodal 3D models to support open-vocabulary point cloud recognition. Notably, our solution operates with efficiency comparable to zero-shot inference, as it is entirely training-free. Point-Cache demonstrates substantial gains across 8 challenging benchmarks and 4 representative large 3D models, highlighting its effectiveness.

****

Think Small, Act Big: Primitive Prompt Learning for Lifelong Robot Manipulation

Yuanqi Yao, Siao Liu, Haoming Song, Delin Qu, Qizhi Chen, Yan Ding, Bin Zhao, Zh
igang Wang, Xuelong Li, Dong Wang; Proceedings of the Computer Vision and Patter
n Recognition Conference (CVPR), 2025, pp. 22573-22583
Learning a generalist robot that can effectively leverage prior knowledge for co
ntinuous skill acquisition remains significantly challenging. Despite the succes
s of experience replay and parameter-efficient methods in maintaining knowledge
across skills, naively applying these methods causes a failure to leverage the s
hared primitives between skills. To tackle these issues, we propose Primitive Pr
ompt Learning (PPL), to achieve lifelong robot manipulation via reusable and ext
ensible primitives. Within our two stage learning scheme, we first learn a set o
f primitive prompts to model primitives through multi-skills pre-training stage,
 where motion-aware prompts are learned to capture semantic and motion shared pr
imitives across different skills. Secondly, when acquiring new skills in lifelon
g span, new prompts are concatenated and optimized with frozen pretrained prompt
s, boosting the learning via knowledge transfer from old skills to new ones. For
 evaluation, we construct a large-scale skill dataset and conduct extensive expe
riments in both simulation and real-world tasks, demonstrating PPL's superior pe
rformance over state-of-the-art methods.
****
Stop Walking in Circles! Bailing Out Early in Projected Gradient Descent
Philip Doldo, Derek Everett, Amol Khanna, Andre T Nguyen, Edward Raff; Proceedin
gs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6
373-6382
Projected Gradient Descent (PGD) under the L_infinity ball has become one of the
 defacto methods used in adversarial robustness evaluation for computer vision (
CV) due to its reliability and efficacy, making a strong and easy-to-implement i
terative baseline. However, PGD is computationally demanding to apply, especiall
y when thousands of iterations is current best-practice recommendations to gener
ate an adversarial example for a single image. In this work, we introduce a simp
le novel method for early termination of PGD based on cycle detection by exploit
ing the geometry of how PGD is implemented in practice and show that it can prod
uce large speedup factors while providing the exact same estimate of model robus
tness as standard PGD. This method substantially speeds up PGD without sacrifici
ng any attack strength, enabling evaluations of robustness that were previously
computationally intractable
****
MoManipVLA: Transferring Vision-language-action Models for General Mobile Manipu
lation
Zhenyu Wu, Yuheng Zhou, Xiuwei Xu, Ziwei Wang, Haibin Yan; Proceedings of the Co
mputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1714-1723
Mobile manipulation is the fundamental challenge for robotics to assist humans w
ith diverse tasks and environments in everyday life. However, conventional mobil
e manipulation approaches often struggle to generalize across different tasks an
d environments because of the lack of large-scale training.In contrast, recent a
dvances in vision-language-action (VLA) models have shown impressive generalizat
ion capabilities, but these foundation models are developed for fixed-base manip
ulation tasks.Therefore, we propose an efficient policy adaptation framework to
transfer pre-trained VLA models of fix-base manipulation to mobile manipulation,
 so that high generalization ability across tasks and environments can be achiev
ed in mobile manipulation policy.Specifically, we utilize pre-trained VLA models
 to generate waypoints of the end-effector with high generalization ability. We
design motion planning objectives for the mobile base and the robot arm, which a
im at maximizing the physical feasibility of the trajectory. Finally, we present
 an efficient bi-level objective optimization framework for trajectory generatio
n, where the upper-level optimization predicts waypoints for base movement to en
hance the manipulator policy space, and the lower-level optimization selects the
 optimal end-effector trajectory to complete the manipulation task. Extensive ex
perimental results on OVMM and the real world demonstrate that our method achiev
es a 4.2% higher success rate than the state-of-the-art mobile manipulation, and
 only requires 50 training cost for real world deployment due to the strong gene

ralization ability in the pre-trained VLA models.
****
SoMA: Singular Value Decomposed Minor Components Adaptation for Domain Generaliz able Representation Learning
Seokju Yun, Seunghye Chae, Dongheon Lee, Youngmin Ro; Proceedings of the Compute r Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25602-25612
Domain generalization (DG) aims to adapt a model using one or multiple source do mains to ensure robust performance in unseen target domains. Recently, Parameter -Efficient Fine-Tuning (PEFT) of foundation models has shown promising results i n the context of DG problem. Nevertheless, existing PEFT methods still struggle to strike a balance between preserving generalizable components of the pre-train ed model and learning task-specific features. To gain insights into the distribu tion of generalizable components, we begin by analyzing the pre-trained weights through the lens of singular value decomposition. Building on these insights, w e introduce Singular Value Decomposed Minor Components Adaptation (SoMA), an app roach that selectively tunes minor singular components while keeping the residua l parts frozen. SoMA effectively retains the generalization ability of the pre-t rained model while efficiently acquiring task-specific skills. Moreover, we free ze domain-generalizable blocks and employ an annealing weight decay strategy, th ereby achieving an optimal balance in the delicate trade-off between generalizab ility and discriminability. SoMA attains state-of-the-art results on multiple be nchmarks that span both domain generalized semantic segmentation to domain gener alized object detection. In addition, our methods introduce no additional infere nce overhead or regularization loss, maintain compatibility with any backbone or head, and are designed to be versatile, allowing easy integration into a wide r ange of tasks.
****
Depth-Guided Bundle Sampling for Efficient Generalizable Neural Radiance Field R econstruction
Li Fang, Hao Zhu, Longlong Chen, Fei Hu, Long Ye, Zhan Ma; Proceedings of the Co mputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11217-11226
Recent advancements in generalizable novel view synthesis have achieved impressi ve quality through interpolation between nearby views. However, rendering high-r esolution images remains computationally intensive due to the need for dense sam pling of all rays. Recognizing that natural scenes are typically piecewise smoot h and sampling all rays is often redundant, we propose a novel depth-guided bund le sampling strategy to accelerate rendering. By grouping adjacent rays into a b undle and sampling them collectively, a shared representation is generated for d ecoding all rays within the bundle. To further optimize efficiency, our adaptive sampling strategy dynamically allocates samples based on depth confidence, conc entrating more samples in complex regions while reducing them in smoother areas. When applied to ENeRF, our method achieves up to a 1.27 dB PSNR improvement and a 47% increase in FPS on the DTU dataset. Extensive experiments on synthetic an d real-world datasets demonstrate state-of-the-art rendering quality and up to 2 x faster rendering compared to existing generalizable methods. Code is available at https://github.com/KLMAV-CUC/GDB-NeRF.
****
TinyFusion: Diffusion Transformers Learned Shallow
Gongfan Fang, Kunjun Li, Xinyin Ma, Xinchao Wang; Proceedings of the Computer Vi sion and Pattern Recognition Conference (CVPR), 2025, pp. 18144-18154
Diffusion Transformers have demonstrated remarkable capabilities in image genera tion but often come with excessive parameterization, resulting in considerable i nference overhead in real-world applications. In this work, we present TinyFusio n, a depth pruning method designed to remove redundant layers from diffusion tra nsformers via end-to-end learning. The core principle of our approach is to crea te a pruned model with high recoverability, allowing it to regain strong perform ance after fine-tuning. To accomplish this, we introduce a differentiable sampli ng technique to make pruning learnable, paired with a co-optimized parameter to simulate future fine-tuning. While prior works focus on minimizing loss or error after pruning, our method explicitly models and optimizes the post-fine-tuning

performance of pruned models. Experimental results indicate that this learnable paradigm offers substantial benefits for layer pruning of diffusion transformers, surpassing existing importance-based and error-based methods. Additionally, TinyFusion exhibits strong generalization across diverse architectures, such as DiTs, MARs, and SiTs. Experiments with DiT-XL show that TinyFusion can craft a shallow diffusion transformer at less than 7% of the pre-training cost, achieving a 2xspeedup with an FID score of 2.86, outperforming competitors with comparable efficiency. Code is available at https://github.com/VainF/TinyFusion

****

Ref-GS: Directional Factorization for 2D Gaussian Splatting

Youjia Zhang, Anpei Chen, Yumin Wan, Zikai Song, Junqing Yu, Yawei Luo, Wei Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26483-26492

In this paper, we introduce Ref-GS, a novel approach for directional light factorization in 2D Gaussian splatting, which enables photorealistic view-dependent appearance rendering and precise geometry recovery. Ref-GS builds upon the deferred rendering of Gaussian splatting and applies directional encoding to the deferred-rendered surface, effectively reducing the ambiguity between orientation and viewing angle. Next, we introduce a spherical mip-grid to capture varying levels of surface roughness, enabling roughness-aware Gaussian shading. Additionally, we propose a simple yet efficient geometry-lighting factorization that connects geometry and lighting via the vector outer product, significantly reducing renderer overhead when integrating volumetric attributes. Our method achieves superior photorealistic rendering for a range of open-world scenes while also accurately recovering geometry.

****

SVG-IR: Spatially-Varying Gaussian Splatting for Inverse Rendering

Hanxiao Sun, Yupeng Gao, Jin Xie, Jian Yang, Beibei Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16143-16152

Reconstructing 3D assets from images, known as inverse rendering (IR), remains a challenging task due to its ill-posed nature. 3D Gaussian Splatting (3DGS) has demonstrated impressive capabilities for novel view synthesis (NVS) tasks. Methods apply it to relighting by separating radiance into BRDF parameters and lighting, yet produce inferior relighting quality with artifacts and unnatural indirect illumination due to the limited capability of each Gaussian, which has constant material parameters and normal, alongside the absence of physical constraints for indirect lighting. In this paper, we present a novel framework called Spatially-vayring Gaussian Inverse Rendering (SVG-IR), aimed at enhancing both NVS and relighting quality. To this end, we propose a new representation--Spatially-varying Gaussian (SVG)--that allows per-Gaussian spatially varying parameters. This enhanced representation is complemented by a SVG splatting scheme akin to vertex/fragment shading in traditional graphics pipelines. Furthermore, we integrate a physically-based indirect lighting model, enabling more realistic relighting. The proposed SVG-IR framework significantly improves rendering quality, outperforming state-of-the-art NeRF-based methods by 2.5 dB in peak signal-to-noise ratio (PSNR) and surpassing existing Gaussian-based techniques by 3.5 dB in relighting tasks, all while maintaining a real-time rendering speed. The source code is available at https://github.com/learner-shx/SVG-IR.

****

Stable-SCore: A Stable Registration-based Framework for 3D Shape Correspondence

Haolin Liu, Xiaohang Zhan, Zizheng Yan, Zhongjin Luo, Yuxin Wen, Xiaoguang Han; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 917-928

Establishing character shape correspondence is a critical and fundamental task in computer vision and graphics, with diverse applications including re-topology, attribute transfer, and shape interpolation. Current dominant functional map methods, while effective in controlled scenarios, struggle in real situations with more complex challenges such as non-isometric shape discrepancies. In response, we revisit registration-for-correspondence methods and tap their potential for more stable shape correspondence estimation. To overcome their common issues inc

luding unstable deformations and the necessity for careful pre-alignment or high-quality initial 3D correspondences, we introduce Stable-SCore: A Stable Registration-based Framework for 3D Shape Correspondence. We first re-purpose a foundation model for 2D character correspondence that ensures reliable and stable 2D mappings. Crucially, we propose a novel Semantic Flow Guided Registration approach that leverages 2D correspondence to guide mesh deformations. Our framework significantly surpasses existing methods in challenging scenarios, and brings possibilities for a wide array of real applications, as demonstrated in our results.
****

Beyond Single-Modal Boundary: Cross-Modal Anomaly Detection through Visual Prototype and Harmonization

Kai Mao, Ping Wei, Yiyang Lian, Yangyang Wang, Nanning Zheng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9964-9973

Anomaly detection is a significant task for its application and research value. While existing methods have made impressive progress within the same modality, cross-modal anomaly detection remains an open and challenging problem. In this paper, we propose a cross-modal anomaly detection model that is trained using data from a variety of existing modalities and can be generalized well to unseen modalities. The model consists of three major components: 1) the Transferable Visual Prototype directly learns normal/abnormal semantics in visual space; 2) the Prototype Harmonization strategy adaptively utilizes the Transferable Visual Prototypes from various modalities for inference on the unknown modality; 3) the Visual Discrepancy Inference under the few-shot setting enhances performance. In the zero-shot setting, the proposed method achieves AUROC improvements of 4.1%, 6.1%, 7.6%, and 6.8% over the best competing methods in the RGB, 3D, MRI/CT, and Thermal modalities, respectively. In the few-shot setting, our model also achieves the highest AUROC/AP on ten datasets in four modalities, substantially outperforming existing methods. Codes are available at https://github.com/Kerio99/CMAD.
****

VDocRAG: Retrieval-Augmented Generation over Visually-Rich Documents

Ryota Tanaka, Taichi Iki, Taku Hasegawa, Kyosuke Nishida, Kuniko Saito, Jun Suzuki; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24827-24837

We aim to develop a retrieval-augmented generation (RAG) framework that answers questions over a corpus of visually-rich documents presented in mixed modalities (e.g., charts, tables) and diverse formats (e.g., PDF, PPTX). In this paper, we introduce a new RAG framework, VDocRAG, which can directly understand varied documents and modalities in a unified image format to prevent missing information that occurs by parsing documents to obtain text. To improve the performance, we propose novel self-supervised pre-training tasks that adapt large vision-language models for retrieval by compressing visual information into dense token representations while aligning them with textual content in documents. Furthermore, we introduce OpenDocVQA, the first unified collection of open-domain document visual question answering datasets, encompassing diverse document types and formats. OpenDocVQA provides a comprehensive resource for training and evaluating retrieval and question answering models on visually-rich documents in an open-domain setting. Experiments show that VDocRAG substantially outperforms conventional text-based RAG and has strong generalization capability, highlighting the potential of an effective RAG paradigm for real-world documents.
****

Align-KD: Distilling Cross-Modal Alignment Knowledge for Mobile Vision-Language Large Model Enhancement

Qianhan Feng, Wenshuo Li, Tong Lin, Xinghao Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4178-4188

Vision-Language Models (VLMs) bring powerful understanding and reasoning capabilities to multimodal tasks. Meanwhile, the great need for capable aritificial intelligence on mobile devices also arises, such as the AI assistant software. Some efforts try to migrate VLMs to edge devices to expand their application scope. Simplifying the model structure is a common method, but as the model shrinks, the trade-off between performance and size becomes more and more difficult. Knowle

dge distillation (KD) can help models improve comprehensive capabilities without increasing size or data volume. However, most of the existing large model distillation techniques only consider applications on single-modal LLMs, or only use teachers to create new data environments for students. None of these methods take into account the distillation of the most important cross-modal alignment knowledge in VLMs. We propose a method called Align-KD to guide the student model to learn the cross-modal matching that occurs at the shallow layer. The teacher also helps student learn the projection of vision token into text embedding space based on the focus of text. Under the guidance of Align-KD, the 1.7B MobileVLM V2 model can learn rich knowledge from the 7B teacher model with light design of training loss, and achieve an average score improvement of 2.0 across 6 benchmarks under two training subsets respectively. Code is available at: https://github.com/fqhank/Align-KD.

****

Pose Priors from Language Models

Sanjay Subramanian, Evonne Ng, Lea Müller, Dan Klein, Shiry Ginosar, Trevor Darrell; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7125-7135

Language is often used to describe physical interaction, yet most 3D human pose estimation methods overlook this rich source of information. We bridge this gap by leveraging large multimodal models (LMMs) as priors for reconstructing contact poses, offering a scalable alternative to traditional methods that rely on human annotations or motion capture data. Our approach extracts contact-relevant descriptors from an LMM and translates them into tractable losses to constrain 3D human pose optimization. Despite its simplicity, our method produces compelling reconstructions for both two-person interactions and self-contact scenarios, accurately capturing the semantics of physical and social interactions. Our results demonstrate that LMMs can serve as powerful tools for contact prediction and pose estimation, offering an alternative to costly manual human annotations or motion capture data. Our code is publicly available at prosepose.github.io.

****

Concept Lancet: Image Editing with Compositional Representation Transplant

Jinqi Luo, Tianjiao Ding, Kwan Ho Ryan Chan, Hancheng Min, Chris Callison-Burch, Rene Vidal; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28502-28512

Diffusion models are widely used for image editing tasks. Existing editing methods often design a representation manipulation procedure by curating an edit direction in the text embedding or score space. However, such a procedure faces a key challenge: overestimating the edit strength harms visual consistency while underestimating it fails the editing task. Notably, each source image may require a different editing strength, and it is costly to search for an appropriate strength via trial-and-error. To address this challenge, we propose Concept Lancet (CoLan), a zero-shot plug-and-play framework for principled representation manipulation in diffusion-based image editing. At inference time, we decompose the source input in the latent (text embedding or diffusion score) space as a sparse linear combination of the representations of the collected visual concepts and phrases. This allows us to accurately estimate the presence of concepts in each image, which informs the edit. Based on the editing task (replace/add/remove), we perform a customized concept transplant process to impose the corresponding editing direction. To sufficiently model the concept space, we curate a conceptual representation dataset, CoLan-150K, which contains diverse descriptions and scenarios of visual concepts and phrases for the latent dictionary. Experiments on multiple diffusion-based image editing baselines show that methods equipped with CoLan achieve state-of-the-art performance in editing effectiveness and consistency preservation.

****

Scaling Mesh Generation via Compressive Tokenization

Haohan Weng, Zibo Zhao, Biwen Lei, Xianghui Yang, Jian Liu, Zeqiang Lai, Zhuo Chen, Yuhong Liu, Jie Jiang, Chunchao Guo, Tong Zhang, Shenghua Gao, C.L. Philip Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR

), 2025, pp. 11093-11103
We propose a compressive yet effective mesh tokenization, Blocked and Patchified Tokenization (BPT), facilitating the generation of meshes exceeding 8k faces. BPT compresses mesh sequences by employing block-wise indexing and patch aggregation, reducing their length by approximately 75% compared to the vanilla sequences. This compression milestone unlocks the potential to utilize mesh data with significantly more faces, thereby enhancing detail richness and improving generation robustness. Empowered with the BPT, we have built a foundation mesh generative model training on scaled mesh data to support flexible control for point clouds and images. Our model demonstrates the capability to generate meshes with intricate details and accurate topology, achieving SoTA performance on mesh generation and reaching the level for direct product usage.
****

Generative Densification: Learning to Densify Gaussians for High-Fidelity Generalizable 3D Reconstruction
Seungtae Nam, Xiangyu Sun, Gyeongjin Kang, Younggeun Lee, Seungjun Oh, Eunbyung Park; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26683-26693
Generalized feed-forward Gaussian models have made significant strides in sparse-view 3D reconstruction by leveraging prior knowledge from large multi-view datasets. However, these models often struggle to represent high-frequency details primarily due to the limited number of Gaussians. While the densification strategy from per-scene 3D Gaussian splatting can be adapted for feed-forward models, it typically requires tens of thousands of optimization steps to reconstruct fine details and can easily lead to overfitting in sparse-view scenarios. In this paper, we propose Generative Densification, an efficient and generalizable densification strategy specifically tailored for feed-forward models. Instead of iteratively splitting and cloning raw Gaussian parameters, our method up-samples feature representations produced by feed-forward models and generates their corresponding fine Gaussians in a single forward pass, leveraging learned prior knowledge for enhanced generalization. Experimental results on both object-level and scene-level reconstruction tasks demonstrate that our method outperforms state-of-the-art approaches with comparable or smaller model sizes, achieving notable improvements in representing fine details.
****

LogoSP: Local-global Grouping of Superpoints for Unsupervised Semantic Segmentation of 3D Point Clouds
Zihui Zhang, Weisheng Dai, Hongtao Wen, Bo Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1374-1384
We study the problem of unsupervised 3D semantic segmentation on raw point clouds without needing human labels in training. Existing methods usually formulate this problem into learning per-point local features followed by a simple grouping strategy, lacking the ability to discover additional and possibly richer semantic priors beyond local features. In this paper, we introduce LogoSP to learn 3D semantics from both local and global point features. The key to our approach is to discover 3D semantic information by grouping superpoints according to their global patterns in the frequency domain, thus generating highly accurate semantic pseudo-labels for training a segmentation network. Extensive experiments on two indoor and an outdoor datasets show that our LogoSP surpasses all existing unsupervised methods by large margins, achieving the state-of-the-art performance for unsupervised 3D semantic segmentation. Notably, our investigation into the learned global patterns reveals that they truly represent meaningful 3D semantics in the absence of human labels during training. Our code and data are available at https://github.com/vLARgroup/LogoSP
****

Exploring Intrinsic Normal Prototypes within a Single Image for Universal Anomaly Detection
Wei Luo, Yunkang Cao, Haiming Yao, Xiaotian Zhang, Jianan Lou, Yuqi Cheng, Weiming Shen, Wenyong Yu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9974-9983

Anomaly detection (AD) is essential for industrial inspection, yet existing meth
ods typically rely on "comparing" test images to normal references from a traini
ng set. However, variations in appearance and positioning often complicate the a
lignment of these references with the test image, limiting detection accuracy. W
e observe that most anomalies manifest as local variations, meaning that even wi
thin anomalous images, valuable normal information remains. We argue that this i
nformation is useful and may be more aligned with the anomalies since both the a
nomalies and the normal information originate from the same image. Therefore, ra
ther than relying on external normality from the training set, we propose INP-Fo
rmer, a novel method that extracts Intrinsic Normal Prototypes (INPs) directly f
rom the test image. Specifically, we introduce the INP Extractor, which linearly
 combines normal tokens to represent INPs. We further propose an INP Coherence L
oss to ensure INPs can faithfully represent normality for the testing image. The
se INPs then guide the INP-Guided Decoder to reconstruct only normal tokens, wit
h reconstruction errors serving as anomaly scores. Additionally, we propose a So
ft Mining Loss to prioritize hard-to-optimize samples during training. INP-Forme
r achieves state-of-the-art performance in single-class, multi-class, and few-sh
ot AD tasks across MVTec-AD, VisA, and Real-IAD, positioning it as a versatile a
nd universal solution for AD. Remarkably, INP-Former also demonstrates some zero
-shot AD capability. Code is available at: https://github.com/luow23/INP-Former
.
****

Augmenting Perceptual Super-Resolution via Image Quality Predictors
Fengjia Zhang, Samrudhdhi B. Rangrej, Tristan Aumentado-Armstrong, Afsaneh Fazly
, Alex Levinshtein; Proceedings of the Computer Vision and Pattern Recognition C
onference (CVPR), 2025, pp. 2311-2322
Super-resolution (SR), a classical inverse problem in computer vision, is inhere
ntly ill-posed, inducing a distribution of plausible solutions for every input.
However, the desired result is not simply the expectation of this distribution,
which is the blurry image obtained by minimizing pixelwise error, but rather the
 sample with the highest image quality. A variety of techniques, from perceptual
 metrics to adversarial losses, are employed to this end. In this work, we explo
re an alternative: utilizing powerful non-reference image quality assessment (NR
-IQA) models in the SR context. We begin with a comprehensive analysis of NR-IQA
 metrics on human-derived SR data, identifying both the accuracy (human alignmen
t) and complementarity of different metrics. Then, we explore two methods of app
lying NR-IQA models to SR learning: (i) altering data sampling, by building on a
n existing multi-ground-truth SR framework, and (ii) directly optimizing a diffe
rentiable quality score. Our results demonstrate a more human-centric perception
-distortion tradeoff, focusing less on non-perceptual pixelwise distortion, inst
ead improving the balance between perceptual fidelity and human-tuned NR-IQA mea
sures.
****

TurboFill: Adapting Few-step Text-to-image Model for Fast Image Inpainting
Liangbin Xie, Daniil Pakhomov, Zhonghao Wang, Zongze Wu, Ziyan Chen, Yuqian Zhou
, Haitian Zheng, Zhifei Zhang, Zhe Lin, Jiantao Zhou, Chao Dong; Proceedings of
the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7613-76
22
This paper introduces TurboFill, a fast image inpainting model that enhances a f
ew-step text-to-image diffusion model with an inpainting adapter for high-qualit
y and efficient inpainting. While standard diffusion models generate high-qualit
y results, they incur high computational costs. We overcome this by training an
inpainting adapter on a few-step distilled text-to-image model, DMD2, using a no
vel 3-step adversarial training scheme to ensure realistic, structurally consist
ent, and visually harmonious inpainted regions. To evaluate TurboFill, we propos
e two benchmarks: DilationBench, which tests performance across mask sizes, and
HumanBench, based on human feedback for complex prompts. Experiments show that T
urboFill outperforms both multi-step BrushNet and few-step inpainting methods, s
etting a new benchmark for high-performance inpainting tasks. The project page i
s available \href https://liangbinxie.github.io/projects/TurboFill/ here

****
Stochastic Human Motion Prediction with Memory of Action Transition and Action Characteristic
Jianwei Tang, Hong Yang, Tengyue Chen, Jian-Fang Hu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1883-1893
Action-driven stochastic human motion prediction aims to generate future motion sequences of a pre-defined target action based on given past observed sequences performing non-target actions. This task primarily presents two challenges. Firstly, generating smooth transition motions is hard due to the varying transition speeds of different actions. Secondly, the action characteristic is difficult to be learned because of the similarity of some actions. These issues cause the predicted results to be unreasonable and inconsistent. As a result, we propose two memory banks, the Soft-transition Action Bank (STAB) and Action Characteristic Bank (ACB), to tackle the problems above. The STAB stores the action transition information. It is equipped with the novel soft searching approach, which encourages the model to focus on multiple possible action categories of observed motions. The ACB records action characteristic, which produces more prior information for predicting certain actions. To fuse the features retrieved from the two banks better, we further propose the Adaptive Attention Adjustment (AAA) strategy. Extensive experiments on four motion prediction datasets demonstrate that our approach consistently outperforms the previous state-of-the-art. The demo and code are available at https://hyqlat.github.io/STABACB.github.io/.
****
Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis
Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, Xing Sun; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24108-24118
In the quest for artificial general intelligence, Multi-modal Large Language Models (MLLMs) have emerged as a focal point in recent advancements. However, the predominant focus remains on developing their capabilities in static image understanding. The potential of MLLMs to process sequential visual data is still insufficiently explored, highlighting the lack of a comprehensive, high-quality assessment of their performance. In this paper, we introduce Video-MME, the first-ever full-spectrum, Multi-Modal Evaluation benchmark of MLLMs in Video analysis. Our work distinguishes from existing benchmarks through four key features: 1) Diversity in video types, spanning 6 primary visual domains with 30 subfields to ensure broad scenario generalizability; 2) Duration in temporal dimension, encompassing both short-, medium-, and long-term videos, ranging from 11 seconds to 1 hour, for robust contextual dynamics; 3) Breadth in data modalities, integrating multi-modal inputs besides video frames, including subtitles and audios, to unveil the all-round capabilities of MLLMs; 4) Quality in annotations, utilizing rigorous manual labeling by expert annotators to facilitate precise and reliable model assessment. With Video-MME, we extensively evaluate various state-of-the-art MLLMs, and reveal that Gemini 1.5 Pro is the best-performing commercial model, significantly outperforming the open-source models with an average accuracy of 75%, compared to 71.9% for GPT-4o. The results also demonstrate that Video-MME is a universal benchmark that applies to both image and video MLLMs. Further analysis indicates that subtitle and audio information could significantly enhance video understanding. Besides, a decline in MLLM performance is observed as video duration increases for all models. Our dataset along with these findings underscores the need for further improvements in handling longer sequences and multi-modal data, shedding light on future MLLM development.
****
Perception Tokens Enhance Visual Reasoning in Multimodal Language Models
Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G. Shapiro, Ranjay Krishna; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3836-3845

Multimodal language models (MLMs) still face challenges in fundamental visual perception tasks where specialized models excel. Tasks requiring reasoning about 3D structures benefit from depth estimation, and reasoning about 2D object instances benefits from object detection. Yet, MLMs can not produce intermediate depth or boxes to reason over. Finetuning MLMs on relevant data doesn't generalize well and outsourcing computation to specialized vision tools is too compute-intensive and memory-inefficient. To address this, we introduce Perception Tokens, intrinsic image representations designed to assist reasoning tasks where language is insufficient. Perception tokens act as auxiliary reasoning tokens, akin to chain-of-thought prompts in language models. For example, in a depth-related task, an MLM augmented with perception tokens can reason by generating a depth map as tokens, enabling it to solve the problem effectively. We propose AURORA, a training method that augments MLMs with perception tokens for improved reasoning over visual inputs. AUROA leverages a VQVAE to transform intermediate image representations, such as depth maps into a tokenized format and bounding box tokens, which is then used in a multi-task training framework. AURORA achieves notable improvements across counting benchmarks: +10.8% on BLINK, +11.3% on CVBench, and +8.3% on SEED-Bench, outperforming finetuning approaches in generalization across datasets. It also improves on relative depth: over +6% on BLINK. With perception tokens, AURORA expands the scope of MLMs beyond language-based reasoning, paving the way for more effective visual reasoning capabilities. Code and data will be released at the https://aurora-perception.github.io.

****

Are Images Indistinguishable to Humans Also Indistinguishable to Classifiers?

Zebin You, Xinyu Zhang, Hanzhong Guo, Jingdong Wang, Chongxuan Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28790-28800

The ultimate goal of generative models is to perfectly capture the data distribution. For image generation, common metrics of visual quality (e.g., FID) and the perceived truthfulness of generated images seem to suggest that we are nearing this goal. However, through distribution classification tasks, we reveal that, from the perspective of neural network-based classifiers, even advanced diffusion models are still far from this goal. Specifically, classifiers are able to consistently and effortlessly distinguish real images from generated ones across various settings. Moreover, we uncover an intriguing discrepancy: classifiers can easily differentiate between diffusion models with comparable performance (e.g., U-ViT-H vs. DiT-XL), but struggle to distinguish between models within the same family but of different scales (e.g., EDM2-XS vs. EDM2-XXL). Our methodology carries several important implications. First, it naturally serves as a diagnostic tool for diffusion models by analyzing specific features of generated data. Second, it sheds light on the model autophagy disorder and offers insights into the use of generated data: augmenting real data with generated data is more effective than replacing it. Third, classifier guidance can significantly enhance the realism of generated images.

****

X-Dyna: Expressive Dynamic Human Image Animation

Di Chang, Hongyi Xu, You Xie, Yipeng Gao, Zhengfei Kuang, Shengqu Cai, Chenxu Zhang, Guoxian Song, Chao Wang, Yichun Shi, Zeyuan Chen, Shijie Zhou, Linjie Luo, Gordon Wetzstein, Mohammad Soleymani; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5499-5509

We introduce X-Dyna, a novel zero-shot, diffusion-based pipeline for animating a single human image using facial expressions and body movements derived from a driving video, that generates realistic, context-aware dynamics for both the subject and the surrounding environment. Building on prior approaches centered on human pose control, X-Dyna addresses key factors underlying the loss of dynamic details, enhancing the lifelike qualities of human video animations.At the core of our approach is the Dynamics-Adapter, a lightweight module that effectively integrates reference appearance context into the spatial attentions of the diffusion backbone while preserving the capacity of motion modules in synthesizing fluid and intricate dynamic details. Beyond body pose control, we connect a local con

trol module with our model to capture identity-disentangled facial expressions, facilitating accurate expression transfer for enhanced realism in animated scene s. Together, these components form a unified framework capable of learning physi cal human motion and natural scene dynamics from a diverse blend of human and sc ene videos. Comprehensive qualitative and quantitative evaluations demonstrate t hat X-Dyna outperforms state-of-the-art methods, creating highly lifelike and ex pressive animations. The code is available at https://github.com/bytedance/X-Dyn a
****

Understanding Multi-layered Transmission Matrices
Anat Levin, Marina Alterman; Proceedings of the Computer Vision and Pattern Reco gnition Conference (CVPR), 2025, pp. 23164-23173
Transmission matrices, mapping the propagation of light from one end of the tiss ue to the other, form an important mathematical tool in the analysis of tissue s cattering and the design of wavefront shaping systems. To understand the relatio nship between their content and the volumetric structure of the tissue, we wish to fit them with multi-slice models, composed of a set of planar aberrations spa ced throughout the volume. The number of layers used in such a model would large ly affect the amount of information compression and the ease in which we can use such layered models in a wavefront-shaping system. This work offers a theoretic al study of such multi-layered models. We attempt to understand how many layers are required for a good fit, and how does the approximation degrade when a small er number of such layers is used. We show analytically that transmission matric es can be well fitted with very sparse layers. This leads to optimistic predicti ons on our ability to use them to design future wavefront shaping systems which can correct tissue aberration over a wide field-of-view.
****

GS-DiT: Advancing Video Generation with Dynamic 3D Gaussian Fields through Effic ient Dense 3D Point Tracking
Weikang Bian, Zhaoyang Huang, Xiaoyu Shi, Yijin Li, Fu-Yun Wang, Hongsheng Li; P roceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 202 5, pp. 21717-21727
4D video control is essential in video generation as it enables the use of sophi sticated lens techniques, such as multi-camera shooting and dolly zoom, which ar e currently unsupported by existing methods. Training a video Diffusion Transfor mer (DiT) directly to control 4D content requires expensive multi-view videos. I nspired by Monocular Dynamic novel View Synthesis (MDVS) that optimizes a 4D rep resentation and renders videos according to different 4D elements, such as camer a pose and object motion editing, we bring dynamic 3D Gaussian fields to video g eneration. Specifically, we propose a novel framework that constructs dynamic 3D Gaussian fields with dense 3D point tracking and renders the Gaussian field for all video frames. Then we finetune a pretrained DiT to generate videos followin g the guidance of the rendered video, dubbed as GS-DiT. To boost the training of the GS-DiT, we also propose an efficient Dense 3D Point Tracking (D3D-PT) metho d for the dynamic 3D Gaussian field construction. Our D3D-PT outperforms Spatial Tracker, the state-of-the-art sparse 3D point tracking method, in accuracy and a ccelerates the inference speed by two orders of magnitude. During the inference stage, GS-DiT can generate videos with the same dynamic content while adhering t o different camera parameters, addressing a significant limitation of current vi deo generation models. GS-DiT demonstrates strong generalization capabilities an d extends the 4D controllability of Gaussian splatting to video generation beyon d just camera poses. It supports advanced cinematic effects through the manipula tion of the Gaussian field and camera intrinsics, making it a powerful tool for creative video production. Demos are available at https://wkbian.github.io/Proje cts/GS-DiT/.
****

AnyMoLe: Any Character Motion In-betweening Leveraging Video Diffusion Models
Kwan Yun, Seokhyeon Hong, Chaelin Kim, Junyong Noh; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27838-27848
Despite recent advancements in learning-based motion in-betweening, a key limita

tion has been overlooked: the requirement for character-specific datasets. In th
is work, we introduce AnyMoLe, a novel method that addresses this limitation by
leveraging video diffusion models to generate motion in-between frames for arbit
rary characters without external data. Our approach employs a two-stage frame ge
neration process to enhance contextual understanding. Furthermore, to bridge the
 domain gap between real-world and rendered character animations, we introduce I
CAdapt, a fine-tuning technique for video diffusion models. Additionally, we pro
pose a "motion-video mimicking" optimization technique, enabling seamless motion
 generation for characters with arbitrary joint structures using 2D and 3D-aware
 features. AnyMoLe significantly reduces data dependency while generating smooth
 and realistic transitions, making it applicable to a wide range of motion in-be
tweening tasks.
****

Towards Optimizing Large-Scale Multi-Graph Matching in Bioimaging
Max Kahl, Sebastian Stricker, Lisa Hutschenreiter, Florian Bernard, Carsten Roth
er, Bogdan Savchynskyy; Proceedings of the Computer Vision and Pattern Recogniti
on Conference (CVPR), 2025, pp. 11569-11578
Multi-graph matching is an important problem in computer vision. Our task comes
from bioimaging, where a set of 100 3D-microscopic images of worms have to be br
ought into correspondence. Surprisingly, virtually all existing methods are not
applicable to this large-scale, real-world problem since they either assume a co
mplete or dense problem setting, and they have so far only been applied to small
-scale, toy or synthetic problems. Despite claims in literature that methods add
ressing complete multi-graph matching are applicable in an incomplete setting, o
ur first contribution is to prove that their runtime would be excessive and impr
actical. Our second contribution is a new method for incomplete multi-graph mat
ching that applies to real-world, larger-scale problems. We experimentally show
that for our bioimaging application we are able to attain results in less than t
wo minutes, whereas the only competing approach requires at least half an hour w
hile producing far worse results. Furthermore, even for small-scale, dense or co
mplete problem instances we achieve results that are at least on par with the le
ading methods, but an order of magnitude faster.
****

Towards Effective and Sparse Adversarial Attack on Spiking Neural Networks via B
reaking Invisible Surrogate Gradients
Li Lun, Kunyu Feng, Qinglong Ni, Ling Liang, Yuan Wang, Ying Li, Dunshan Yu, Xia
oxin Cui; Proceedings of the Computer Vision and Pattern Recognition Conference
(CVPR), 2025, pp. 3540-3551
Spiking neural networks (SNNs) have shown their competence in handling spatial-t
emporal event-based data with low energy consumption. Similar to conventional ar
tificial neural networks (ANNs), SNNs are also vulnerable to gradient-based adve
rsarial attacks, wherein gradients are calculated by spatial-temporal back-propa
gation (STBP) and surrogate gradients (SGs). However, the SGs may be invisible f
or an inference-only model as they do not influence the inference results, and c
urrent gradient-based attacks are ineffective for binary dynamic images captured
 by the dynamic vision sensor (DVS). While some approaches addressed the issue o
f invisible SGs through universal SGs, their SGs lack a correlation with the vic
tim model, resulting in sub-optimal performance. Moreover, the imperceptibility
of existing SNN-based binary attacks is still insufficient. In this paper, we in
troduce an innovative potential-dependent surrogate gradient (PDSG) method to es
tablish a robust connection between the SG and the model, thereby enhancing the
adaptability of adversarial attacks across various models with invisible SGs. Ad
ditionally, we propose the sparse dynamic attack (SDA) to effectively attack bin
ary dynamic images. Utilizing a generation-reduction paradigm, SDA can fully opt
imize the sparsity of adversarial perturbations. Experimental results demonstrat
e that our PDSG and SDA outperform state-of-the-art SNN-based attacks across var
ious models and datasets. Specifically, our PDSG achieves 100% attack success ra
te on ImageNet, and our SDA obtains 82% attack success rate by modifying only 0.
24% of the pixels on CIFAR10DVS. The code is available at https://github.com/ryi
me/PDSG-SDA .

****
Towards Understanding and Quantifying Uncertainty for Text-to-Image Generation
Gianni Franchi, Nacim Belkhir, Dat Nguyen Trong, Guoxuan Xia, Andrea Pilzer; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8062-8072
Uncertainty quantification in text-to-image (T2I) generative models is crucial for understanding model behavior and improving output reliability. In this paper, we are the first to quantify and evaluate the uncertainty of T2I models with respect to the prompt. Alongside adapting existing approaches designed to measure uncertainty in the image space, we also introduce Prompt-based UNCertainty Estimation for T2I models (PUNC), a novel method leveraging Large Vision-Language Models (LVLMs) to better address uncertainties arising from the semantics of the prompt and generated images. PUNC utilizes a LVLM to caption a generated image, and then compares the caption with the original prompt in the more semantically meaningful text space. PUNC also enables the disentanglement of both aleatoric and epistemic uncertainties via precision and recall, which image-space approaches are unable to do. Extensive experiments demonstrate that PUNC outperforms state-of-the-art uncertainty estimation techniques across various settings. Uncertainty quantification in text-to-image generation models can be used on various applications including bias detection, copyright protection, and OOD detection. We also introduce a comprehensive dataset of text prompts and generation pairs to foster further research in uncertainty quantification for generative models. Our findings illustrate that PUNC not only achieves competitive performance but also enables novel applications in evaluating and improving the trustworthiness of text-to-image models. The code is available at https://github.com/ENSTA-U2IS-AI/Uncertainty_diffusion
****
PS-Diffusion: Photorealistic Subject-Driven Image Editing with Disentangled Control and Attention
Weicheng Wang, Guoli Jia, Zhongqi Zhang, Liang Lin, Jufeng Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18302-18312
Diffusion models pre-trained on large-scale paired image-text data achieve significant success in image editing. To convey more fine-grained visual details, subject-driven editing integrates subjects in user-provided reference images into existing scenes. However, it is challenging to obtain photorealistic results, which simulate contextual interactions, such as reflections, illumination, and shadows, induced by merging the target object into the source image. To address this issue, we propose PS-Diffusion, which ensures realistic and consistent object-scene blending while maintaining the invariance of subject appearance during editing. To be specific, we first divide the contextual interactions into those occurring in the foreground and the background areas. The effect of the former is estimated through intrinsic image decomposition, and the region of the latter is predicted in an additional background effect control branch. Moreover, we propose an effect attention module to disentangle the learning processes of interaction and subject, alleviating confusion between them. Additionally, we introduce a synthesized dataset, Replace-5K, consisting of 5,000 image pairs with invariant subject and contextual interactions via 3D rendering. Extensive quantitative and qualitative experiments on our dataset and two real-world datasets demonstrate that our method achieves state-of-the-art performance. The code is available in the https://github.com/wei-cheng777/PS-Diffusion.
****
Exploring Visual Vulnerabilities via Multi-Loss Adversarial Search for Jailbreaking Vision-Language Models
Shuyang Hao, Bryan Hooi, Jun Liu, Kai-Wei Chang, Zi Huang, Yujun Cai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19890-19899
Despite inheriting security measures from underlying language models, Vision-Language Models (VLMs) may still be vulnerable to safety alignment issues. Through empirical analysis, we uncover two critical findings: scenario-matched images ca

n significantly amplify harmful outputs, and contrary to common assumptions in g radient-based attacks, minimal loss values do not guarantee optimal attack effec tiveness. Building on these insights, we introduce MLAI (Multi-Loss Adversarial Images), a novel jailbreak framework that leverages scenario-aware image generat ion for semantic alignment, exploits flat minima theory for robust adversarial i mage selection, and employs multi-image collaborative attacks for enhanced effec tiveness. Extensive experiments demonstrate MLAI's significant impact, achieving attack success rates of 77.75% on MiniGPT-4 and 82.80% on LLaVA-2, substantiall y outperforming existing methods by margins of 34.37% and 12.77% respectively. F urthermore, MLAI shows considerable transferability to commercial black-box VLMs , achieving up to 60.11% success rate. Our work reveals fundamental visual vulne rabilities in current VLMs safety mechanisms and underscores the need for strong er defenses. Warning: This paper contains potentially harmful example text.
****

LLaVA-ST: A Multimodal Large Language Model for Fine-Grained Spatial-Temporal Un derstanding

Hongyu Li, Jinyu Chen, Ziyu Wei, Shaofei Huang, Tianrui Hui, Jialin Gao, Xiaomin g Wei, Si Liu; Proceedings of the Computer Vision and Pattern Recognition Confer ence (CVPR), 2025, pp. 8592-8603

Recent advancements in multimodal large language models (MLLMs) have shown promi sing results, yet existing approaches struggle to effectively handle both tempor al and spatial localization simultaneously. This challenge stems from two key is sues: first, incorporating spatial-temporal localization introduces a vast numbe r of coordinate combinations, complicating the alignment of linguistic and visua l coordinate representations; second, encoding fine-grained temporal and spatial information during video feature compression is inherently difficult. To addres s these issues, we propose LLaVA-ST, a MLLM for fine-grained spatial-temporal mu ltimodal understanding. In LLaVA-ST, we propose Language-Aligned Positional Embe dding, which embeds the textual coordinate special token into the visual space, simplifying the alignment of fine-grained spatial-temporal correspondences. Addi tionally, we design the Spatial-Temporal Packer, which decouples the feature com pression of temporal and spatial resolutions into two distinct point-to-region a ttention processing streams. Furthermore, we propose ST-Align dataset with 4.3M training samples for fine-grained spatial-temporal multimodal understanding. Wit h ST-align, we present a progressive training pipeline that aligns the visual an d textual feature through sequential coarse-to-fine stages.Additionally, we intr oduce an ST-Align benchmark to evaluate spatial-temporal interleaved fine-graine d understanding tasks, which include Spatial-Temporal Video Grounding (STVG) , E vent Localization and Captioning (ELC) and Spatial Video Grounding (SVG). LLaVA-ST achieves outstanding performance on 11 benchmarks requiring fine-grained temp oral, spatial, or spatial-temporal interleaving multimodal understanding.
****

Insight-V: Exploring Long-Chain Visual Reasoning with Multimodal Large Language Models

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, Zi wei Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9062-9072

Large Language Models (LLMs) demonstrate enhanced capabilities and reliability b y reasoning more, evolving from Chain-of-Thought prompting to product-level solu tions like OpenAI o1. Despite various efforts to improve LLM reasoning, high-qu ality long-chain reasoning data and optimized training pipelines still remain in adequately explored in vision-language tasks. In this paper, we present Insight-V, an early effort to 1) scalably produce long and robust reasoning data for com plex multi-modal tasks, and 2) an effective training pipeline to enhance the rea soning capabilities of multi-modal large language models (MLLMs). Specifically, to create long and structured reasoning data without human labor, we design a tw o-step pipeline with a progressive strategy to generate sufficiently long and di verse reasoning paths and a multi-granularity assessment method to ensure data q uality. We observe that directly supervising MLLMs with such long and complex re asoning data will not yield ideal reasoning ability. To tackle this problem, we

design a multi-agent system consisting of a reasoning agent dedicated to perform ing long-chain reasoning and a summary agent trained to judge and summarize reas oning results. We further incorporate an iterative DPO algorithm to enhance the reasoning agent's generation stability and quality. Based on the popular LLaVA-N eXT model, our method shows an average improvement of 7.5% across seven challeng ing multi-modal benchmarks requiring visual reasoning. We also achieve a 4.2% i mprovement on a stronger base MLLM, highlighting the potential to further advanc e state-of-the-art models. Benefiting from our multi-agent system, Insight-V can also easily maintain or improve performance on perception-focused multi-modal t asks. We will make our data and code publicly available to promote future resear ch in this emerging field.
****

MaIR: A Locality- and Continuity-Preserving Mamba for Image Restoration
Boyun Li, Haiyu Zhao, Wenxin Wang, Peng Hu, Yuanbiao Gou, Xi Peng; Proceedings o f the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7491-7501
Recent advancements in Mamba have shown promising results in image restoration. These methods typically flatten 2D images into multiple distinct 1D sequences al ong rows and columns, process each sequence independently using selective scan o peration, and recombine them to form the outputs. However, such a paradigm overl ooks two vital aspects: i) the local relationships and spatial continuity inhere nt in natural images, and ii) the discrepancies among sequences unfolded through totally different ways. To overcome the drawbacks, we explore two problems in M amba-based restoration methods: i) how to design a scanning strategy preserving both locality and continuity while facilitating restoration, and ii) how to aggr egate the distinct sequences unfolded in totally different ways. To address thes e problems, we propose a novel Mamba-based Image Restoration model (MaIR), which consists of Nested S-shaped Scanning strategy (NSS) and Sequence Shuffle Attent ion block (SSA). Specifically, NSS preserves locality and continuity of the inpu t images through the stripe-based scanning region and the S-shaped scanning path , respectively. SSA aggregates sequences through calculating attention weights w ithin the corresponding channels of different sequences. Thanks to NSS and SSA, MaIR surpasses 40 baselines across 14 challenging datasets, achieving state-of-t he-art performance on the tasks of image super-resolution, denoising, deblurring and dehazing.
****

Few-shot Implicit Function Generation via Equivariance
Suizhi Huang, Xingyi Yang, Hongtao Lu, Xinchao Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16262-16272
Implicit Neural Representations (INRs) have emerged as a powerful framework for representing continuous signals. However, generating diverse INR weights remains challenging due to limited training data. We introduce Few-shot Implicit Functi on Generation, a new problem setup that aims to generate diverse yet functionall y consistent INR weights from only a few examples. This is challenging because e ven for the same signal, the optimal INRs can vary significantly depending on th eir initializations. To tackle this, we propose EquiGen, a framework that can ge nerate new INRs from limited data. The core idea is that functionally similar ne tworks can be transformed into one another through weight permutations, forming an equivariance group. By projecting these weights into an equivariant latent sp ace, we enable diverse generation within these groups, even with few examples. E quiGen implements this through an equivariant encoder trained via contrastive le arning and smooth augmentation, an equivariance-guided diffusion process, and co ntrolled perturbations in the equivariant subspace. Experiments on 2D image and 3D shape INR datasets demonstrate that our approach effectively generates divers e INR weights while preserving their functional properties in few-shot scenarios . Code is available at https://github.com/JeanDiable/EquiGen.
****

RSAR: Restricted State Angle Resolver and Rotated SAR Benchmark
Xin Zhang, Xue Yang, Yuxuan Li, Jian Yang, Ming-Ming Cheng, Xiang Li; Proceeding s of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 74

16-7426
Rotated object detection has made significant progress in the optical remote sen sing. However, advancements in the Synthetic Aperture Radar (SAR) field are lagg ard behind, primarily due to the absence of a large-scale dataset. Annotating su ch a dataset is inefficient and costly. A promising solution is to employ a weak ly supervised model (e.g., trained with available horizontal boxes only) to gene rate pseudo-rotated boxes for reference before manual calibration. Unfortunately , the existing weakly supervised models exhibit limited accuracy in predicting t he object's angle. Previous works attempt to enhance angle prediction by using a ngle resolvers that decouple angles into cosine and sine encodings. In this work , we first reevaluate these resolvers from a unified perspective of dimension ma pping and expose that they share the same shortcomings: these methods overlook t he unit cycle constraint inherent in these encodings, easily leading to predicti on biases. To address this issue, we propose the Unit Cycle Resolver (UCR), whic h incorporates a unit circle constraint loss to improve angle prediction accurac y. Our approach can effectively improve the performance of existing state-of-the -art weakly supervised methods and even surpasses fully supervised models on exi sting optical benchmarks (i.e., DOTA-v1.0). With the aid of UCR, we further anno tate and introduce RSAR, the largest multi-class rotated SAR object detection da taset to date. Extensive experiments on both RSAR and optical datasets demonstra te that our UCR enhances angle prediction accuracy.
****

Dual Energy-Based Model with Open-World Uncertainty Estimation for Out-of-distri bution Detection
Qi Chen, Hu Ding; Proceedings of the Computer Vision and Pattern Recognition Con ference (CVPR), 2025, pp. 25728-25737
Out-of-distribution (OOD) detection is crucial for machine learning models deplo yed in open-world environments. However, existing methods often struggle with mo del over-confidence or rely heavily on empirical energy value estimation, limiti ng their scalability and generalizability. This paper introduces DEBO (Dual Ener gy-Based Model for Out-of-distribution Detection), a novel approach that address es these limitations through an innovative dual classifier architecture and a un ified energy-based objective function. DEBO enhances the standard classification framework by integrating a dual-purpose output space within a single classifier . The primary component classifies in-distribution (ID) data conventionally, whi le the secondary component captures open-world information and estimates uncerta inty. Our method overcomes the dependence of traditional energy model-based OOD detection methods on empirical energy estimation while maintaining theoretical g uarantees. Theoretical analysis demonstrates that DEBO promotes low energy and h igh confidence for ID data, while simultaneously inducing higher energy and decr eased confidence for OOD samples. Extensive experiments conducted on benchmark d atasets reveal that DEBO achieves state-of-the-art OOD detection performance whi le maintaining comparable classification accuracy on ID data.
****

DTGBrepGen: A Novel B-rep Generative Model through Decoupling Topology and Geome try
Jing Li, Yihang Fu, Falai Chen; Proceedings of the Computer Vision and Pattern R ecognition Conference (CVPR), 2025, pp. 21438-21447
Boundary representation (B-rep) of geometric models is a fundamental format in C omputer-Aided Design (CAD). However, automatically generating valid and high-qu ality B-rep models remains challenging due to the complex interdependence betwee n the topology and geometry of the models. Existing methods tend to prioritize g eometric representation while giving insufficient attention to topological const raints, making it difficult to maintain structural validity and geometric accura cy. In this paper, we propose DTGBrepGen, a novel topology-geometry decoupled fr amework for B-rep generation that explicitly addresses both aspects. Our approac h first generates valid topological structures through a two-stage process that independently models edge-face and edge-vertex adjacency relationships. Subseque ntly, we employ Transformer-based diffusion models for sequential geometry gener ation, progressively generating vertex coordinates, followed by edge geometries

and face geometries which are represented as B-splines. Extensive experiments on diverse CAD datasets show that DTGBrepGen significantly outperforms existing methods in both topological validity and geometric accuracy, achieving higher validity rates and producing more diverse and realistic B-reps. Our code is publicly available at https://github.com/jinli99/DTGBrepGen.

****

Continuous Space-Time Video Resampling with Invertible Motion Steganography

Yuantong Zhang, Zhenzhong Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2116-2126

Space-time video resampling aims to conduct both spatial-temporal downsampling and upsampling processes to achieve high-quality video reconstruction.Although there has been much progress, some major challenges still exist, such as how to preserve motion information during temporal resampling while avoiding blurring artifacts, and how to achieve flexible temporal and spatial resampling factors. In this paper, we introduce an Invertible Motion Steganography Module (IMSM), designed to preserve motion information in high-frame-rate videos. This module embeds motion information from high-frame-rate videos into downsampled frames with lower frame rates in a visually imperceptible manner. Its reversible nature allows the motion information to be recovered, facilitating the reconstruction of high-frame-rate videos. Furthermore, we propose a 3D implicit feature modulation technique that enables continuous spatiotemporal resampling. With tailored training strategies, our method supports flexible frame rate conversions, including non-integer changes like 30 FPS to 24 FPS and vice versa. Extensive experiments show that our method significantly outperforms existing solutions across multiple datasets in various video resampling tasks with high flexibility. Codes will be made available.

****

Schedule On the Fly: Diffusion Time Prediction for Faster and Better Image Generation

Zilyu Ye, Zhiyang Chen, Tiancheng Li, Zemin Huang, Weijian Luo, Guo-Jun Qi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23412-23422

Diffusion and flow matching models have achieved remarkable success in text-to-image generation. However, these models typically rely on the predetermined denoising schedules for all prompts. The multi-step reverse diffusion process can be regarded as a kind of chain-of-thought for generating high-quality images step by step. Therefore, diffusion models should reason for each instance to determine the optimal noise schedule adaptively, achieving high generation quality with sampling efficiency. In this paper, we introduce the Time Prediction Diffusion Model (TPDM) for this. TPDM employs a plug-and-play Time Prediction Module (TPM) that predicts the next noise level based on current latent features at each denoising step. We train the TPM using reinforcement learning to maximize a reward that encourages high final image quality while penalizing excessive denoising steps. With such an adaptive scheduler, TPDM not only generates high-quality images that are aligned closely with human preferences but also adjusts diffusion time and the number of denoising steps on the fly, enhancing both performance and efficiency. With Stable Diffusion 3 Medium architecture, TPDM achieves an aesthetic score of 5.44 and a human preference score (HPS) of 29.59, while using around 50% fewer denoising steps to achieve better performance.

****

Learning Audio-guided Video Representation with Gated Attention for Video-Text Retrieval

Boseung Jeong, Jicheol Park, Sungyeon Kim, Suha Kwak; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26202-26211

Video-text retrieval, the task of retrieving videos based on a textual query or vice versa, is of paramount importance for video understanding and multimodal information retrieval. Recent methods in this area rely primarily on visual and textual features and often ignore audio, although it helps enhance overall comprehension of video content.Moreover, traditional models that incorporate audio blindly utilize the audio input regardless of whether it is useful or not, resulting

in suboptimal video representation. To address these limitations, we propose a novel video-text retrieval framework, Audio-guided VIdeo representation learning with GATEd attention (AVIGATE), that effectively leverages audio cues through a gated attention mechanism that selectively filters out uninformative audio signals.In addition, we propose an adaptive margin-based contrastive loss to deal with the inherently unclear positive-negative relationship between video and text, which facilitates learning better video-text alignment.Our extensive experiments demonstrate that AVIGATE achieves state-of-the-art performance on all the public benchmarks.

****

ProtoDepth: Unsupervised Continual Depth Completion with Prototypes
Patrick Rim, Hyoungseob Park, S. Gangopadhyay, Ziyao Zeng, Younjoon Chung, Alex Wong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6304-6316

We present ProtoDepth, a novel prototype-based approach for continual learning of unsupervised depth completion, the multimodal 3D reconstruction task of predicting dense depth maps from RGB images and sparse point clouds. The unsupervised learning paradigm is well-suited for continual learning, as ground truth is not needed. However, when training on new non-stationary distributions, depth completion models will catastrophically forget previously learned information. We address forgetting by learning prototype sets that adapt the latent features of a frozen pretrained model to new domains. Since the original weights are not modified, ProtoDepth does not forget when test-time domain identity is known. To extend ProtoDepth to the challenging setting where the test-time domain identity is withheld, we propose to learn domain descriptors that enable the model to select the appropriate prototype set for inference. We evaluate ProtoDepth on benchmark dataset sequences, where we reduce forgetting compared to baselines by 52.2% for indoor and 53.2% for outdoor to achieve the state of the art.

****

vesselFM: A Foundation Model for Universal 3D Blood Vessel Segmentation
Bastian Wittmann, Yannick Wattenberg, Tamaz Amiranashvili, Suprosanna Shit, Bjoern Menze; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20874-20884

Segmenting 3D blood vessels is a critical yet challenging task in medical image analysis. This is due to significant imaging modality-specific variations in artifacts, vascular patterns and scales, signal-to-noise ratios, and background tissues. These variations, along with domain gaps arising from varying imaging protocols, limit the generalization of existing supervised learning-based methods, requiring tedious voxel-level annotations for each dataset separately. While foundation models promise to alleviate this limitation, they typically fail to generalize to the task of blood vessel segmentation, posing a unique, complex problem. In this work, we present vesselFM, a foundation model designed specifically for the broad task of 3D blood vessel segmentation. Unlike previous models, vesselFM can effortlessly generalize to unseen domains. To achieve zero-shot generalization, we train vesselFM on three heterogeneous data sources: a large, curated annotated dataset, data generated by a domain randomization scheme, and data sampled from a flow matching-based generative model. Extensive evaluations show that vesselFM outperforms state-of-the-art medical image segmentation foundation models across four (pre-)clinically relevant imaging modalities in zero-, one-, and few-shot scenarios, therefore providing a universal solution for 3D blood vessel segmentation.

****

TASTE-Rob: Advancing Video Generation of Task-Oriented Hand-Object Interaction for Generalizable Robotic Manipulation
Hongxiang Zhao, Xingchen Liu, Mutian Xu, Yiming Hao, Weikai Chen, Xiaoguang Han; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27683-27693

We address key limitations in existing datasets and models for task-oriented hand-object interaction video generation, a critical approach of generating video demonstrations for robotic imitation learning. Current datasets, such as Ego4D, o

ften suffer from inconsistent view perspectives and misaligned interactions, lea
ding to reduced video quality and limiting their applicability for precise imita
tion learning tasks. Towards this end, we introduce TASTE-Rob -- a pioneering la
rge-scale dataset of 100,856 ego-centric hand-object interaction videos. Each vi
deo is meticulously aligned with language instructions and recorded from a consi
stent camera viewpoint to ensure interaction clarity. By fine-tuning a Video Dif
fusion Model (VDM) on TASTE-Rob, we achieve realistic object interactions, thoug
h we observed occasional inconsistencies in hand grasping postures. To enhance r
ealism, we introduce a three-stage pose-refinement pipeline that improves hand p
osture accuracy in generated videos. Our curated dataset, coupled with the speci
alized pose-refinement framework, provides notable performance gains in generat
ing high-quality, task-oriented hand-object interaction videos, resulting in ach
ieving superior generalizable robotic manipulation. To foster further advancemen
ts in the field, TASTE-Rob dataset and source code will be made publicly availab
le on our website https://taste-rob.github.io.
****

NoT: Federated Unlearning via Weight Negation
Yasser H. Khalil, Leo Brunswic, Soufiane Lamghari, Xu Li, Mahdi Beitollahi, Xi C
hen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR
), 2025, pp. 25759-25769
Federated unlearning (FU) aims to remove a participant's data contributions from
 a trained federated learning (FL) model, ensuring privacy and regulatory compli
ance. Traditional FU methods often depend on auxiliary storage on either the cli
ent or server side or require direct access to the data targeted for removal--a
dependency that may not be feasible if the data is no longer available. To overc
ome these limitations, we propose NoT, a novel and efficient FU algorithm based
on weight negation (multiplying by -1), which circumvents the need for additiona
l storage and access to the target data. We argue that effective and efficient u
nlearning can be achieved by perturbing model parameters away from the set of op
timal parameters, yet being well-positioned for quick re-optimization. This tech
nique, though seemingly contradictory, is theoretically grounded: we prove that
the weight negation perturbation effectively disrupts inter-layer co-adaptation,
 inducing unlearning while preserving an approximate optimality property, thereb
y enabling rapid recovery. Experimental results across three datasets and three
model architectures demonstrate that NoT significantly outperforms existing base
lines in unlearning efficacy as well as in communication and computational effic
iency.
****

ParaHome: Parameterizing Everyday Home Activities Towards 3D Generative Modeling
 of Human-Object Interactions
Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, Hanbyul Joo; Proceedings of the Compute
r Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1816-1828
To enable machines to understand the way humans interact with the physical world
 in daily life, 3D interaction signals should be captured in natural settings, a
llowing people to engage with multiple objects in a range of sequential and casu
al manipulations. To achieve this goal, we introduce our ParaHome system designe
d to capture dynamic 3D movements of humans and objects within a common home env
ironment. Our system features a multi-view setup with 70 synchronized RGB camer
as, along with wearable motion capture devices including an IMU-based body suit
and hand motion capture gloves. By leveraging the ParaHome system, we collect a
new human-object interaction dataset, including 486 minutes of sequences across
207 captures with 38 participants, offering advancements with three key aspects:
 (1) capturing body motion and dexterous hand manipulation motion alongside mult
iple objects within a contextual home environment; (2) encompassing sequential a
nd concurrent manipulations paired with text descriptions; and (3) including art
iculated objects with multiple parts represented by 3D parameterized models. We
present detailed design justifications for our system, and perform key generativ
e modeling experiments to demonstrate the potential of our dataset.
****

Adapting to the Unknown: Training-Free Audio-Visual Event Perception with Dynami

c Thresholds
Eitan Shaar, Ariel Shaulov, Gal Chechik, Lior Wolf; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3142-3151
In the domain of audio-visual event perception, which focuses on the temporal lo calization and classification of events across distinct modalities (audio and vi sual), existing approaches are constrained by the vocabulary available in their training data. This limitation significantly impedes their capacity to generaliz e to novel, unseen event categories. Furthermore, the annotation process for thi s task is labor-intensive, requiring extensive manual labeling across modalities and temporal segments, limiting the scalability of current methods. Current sta te-of-the-art models ignore the shifts in event distributions over time, reducin g their ability to adjust to changing video dynamics. Additionally, previous met hods rely on late fusion to combine audio and visual information. While straight forward, this approach results in a significant loss of multimodal interactions. To address these challenges, we propose Audio-Visual Adaptive Video Analysis (AV2A), a model-agnostic approach that requires no further training and integrate s a score-level fusion technique to retain richer multimodal interactions. AV2A also includes a within-video label shift algorithm, leveraging input video data and predictions from prior frames to dynamically adjust event distributions for subsequent frames. Moreover, we present the first training-free, open-vocabulary baseline for audio-visual event perception, demonstrating that AV2A achieves su bstantial improvements over naive training-free baselines. We demonstrate the ef fectiveness of AV2A on both zero-shot and weakly-supervised state-of-the-art met hods, achieving notable improvements in performance metrics over existing approa ches.
****

OpenHumanVid: A Large-Scale High-Quality Dataset for Enhancing Human-Centric Vid eo Generation
Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, Siyu Zhu; Proceedings of the Computer Vision and P attern Recognition Conference (CVPR), 2025, pp. 7752-7762
Recent advancements in visual generation technologies have markedly increased th e scale and availability of video datasets, which are crucial for training effec tive video generation models. However, a significant lack of high-quality, human -centric video datasets presents a challenge to progress in this field. To bridg e this gap, we introduce OpenHumanVid, a large-scale and high-quality human-cent ric video dataset characterized by precise and detailed captions that encompass both human appearance and motion states, along with supplementary human motion c onditions, including skeleton sequences and speech audio. To validate the effica cy of this dataset and the associated training strategies, we propose an extensi on of existing classical diffusion transformer architectures and conduct further pretraining of our models on the proposed dataset. Our findings yield two criti cal insights: First, the incorporation of a large-scale, high-quality dataset su bstantially enhances evaluation metrics for generated human videos while preserv ing performance in general video generation tasks. Second, the effective alignme nt of text with human appearance, human motion, and facial motion is essential f or producing high-quality video outputs. Based on these insights and correspondi ng methodologies, the straightforward extended network trained on the proposed d ataset demonstrates an obvious improvement in the generation of human-centric vi deos.
****

Classifier-Free Guidance Inside the Attraction Basin May Cause Memorization
Anubhav Jain, Yuya Kobayashi, Takashi Shibuya, Yuhta Takida, Nasir Memon, Julian Togelius, Yuki Mitsufuji; Proceedings of the Computer Vision and Pattern Recogn ition Conference (CVPR), 2025, pp. 12871-12879
Diffusion models are prone to exactly reproduce images from the training data. T his exact reproduction of the training data is concerning as it can lead to copy right infringement and/or leakage of privacy-sensitive information. In this pape r, we present a novel perspective on the memorization phenomenon and propose a s imple yet effective approach to mitigate it. We argue that memorization occurs b

ecause of an attraction basin in the denoising process which steers the diffusio n trajectory towards a memorized image. However, this can be mitigated by guidin g the diffusion trajectory away from the attraction basin by not applying classi fier-free guidance until an ideal transition point occurs from which classifier- free guidance is applied. This leads to the generation of non-memorized images t hat are high in image quality and well-aligned with the conditioning mechanism. To further improve on this, we present a new guidance technique, opposite guidan ce, that escapes the attraction basin sooner in the denoising process. We demons trate the existence of attraction basins in various scenarios in which memorizat ion occurs, and we show that our proposed approach successfully mitigates memori zation.
****

Track Any Anomalous Object:A Granular Video Anomaly Detection Pipeline
Yuzhi Huang, Chenxin Li, Haitao Zhang, Zixu Lin, Yunlong Lin, Hengyu Liu, Wuyang Li, Xinyu Liu, Jiechao Gao, Yue Huang, Xinghao Ding, Yixuan Yuan; Proceedings o f the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8689-8699
Video anomaly detection (VAD) is crucial in scenarios such as surveillance and a utonomous driving, where timely detection of unexpected activities is essential. Albeit existing methods have primarily focused on detecting anomalous objects i n videos--either by identifying anomalous frames or objects--they often neglect finer-grained analysis, such as anomalous pixels, which limits their ability to capture a broader range of anomalies. To address this challenge, we propose an i nnovative VAD framework called \underline T rack Any \underline A nomalous \unde rline O bject (\textit TAO), which introduces a Granular Video Anomaly Detectio n Framework that, for the first time, integrates the detection of multiple fine- grained anomalous objects into a unified framework. Unlike methods that assign a nomaly scores to every pixel at each moment, our approach transforms the problem into pixel-level tracking of anomalous objects. By linking anomaly scores to su bsequent tasks such as image segmentation and video tracking, our method elimina tes the need for threshold selection and achieves more precise anomaly localizat ion, even in long and challenging video sequences. Experiments on extensive data sets demonstrate that \textit TAO achieves state-of-the-art performance, settin g a new progress for VAD by providing a practical, granular, and holistic soluti on.
****

RANGE: Retrieval Augmented Neural Fields for Multi-Resolution Geo-Embeddings
Aayush Dhakal, Srikumar Sastry, Subash Khanal, Adeel Ahmad, Eric Xing, Nathan Ja cobs; Proceedings of the Computer Vision and Pattern Recognition Conference (CVP R), 2025, pp. 24680-24689
The choice of representation for geographic location significantly impacts the a ccuracy of models for a broad range of geospatial tasks, including fine-grained species classification, population density estimation, and biome classification. Recent works like SatCLIP and GeoCLIP learn such representations by contrastive ly aligning geolocation with co-located images. While these methods work excepti onally well, in this paper, we posit that the current training strategies fail t o fully capture the important visual features. We provide an information theoret ic perspective on why the resulting embeddings from these methods discard crucia l visual information that is important for many downstream tasks. To solve this problem, we propose a novel retrieval-augmented strategy called RANGE. We build our method on the intuition that the visual features of a location can be estima ted by combining the visual features from multiple similar-looking locations. We evaluate our method across a wide variety of tasks. Our results show that RANGE outperforms the existing state-of-the-art models with significant margins in mo st tasks. We show gains of up to 13.1% on classification tasks and 0.145 R^2 on regression tasks. All our code and models will be made available on github.
****

Magma: A Foundation Model for Multimodal AI Agents
Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, Yuquan Deng, Jianfeng Gao; Proceedings

of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14203-14214

We present Magma, a foundation model that serves multimodal AI agentic tasks in both the digital and physical worlds. Magma is a significant extension of vision-language (VL) models in that it not only retains the VL understanding ability (verbal intelligence) of the latter, but is also equipped with the ability to ground and act in the visual-spatial world (spatial-temporal intelligence). To endow agentic capabilities for tasks ranging from UI navigation to robot manipulation, Magma is trained on large amounts of heterogeneous datasets that span from images, videos to robotics data, where actionable visual objects (e.g. clickable buttons in GUI) in images are labeled by Set-of-Mark (SoM) for action grounding, and object movements (e.g. trace of human hands or robotic arms) in videos are labeled by Trace-of-Mark (ToM) for action planning. Extensive experiments show that SoM and ToM help bridge the gap between verbal and action abilities and significantly enhance spatio-temporal intelligence which is fundamental to agentic tasks, as shown in Fig.1. In particular, Magma creates new state-of-the-art results on UI navigation and robotic manipulation tasks, outperforming previous models that are specifically tailored to these tasks. Moreover, Magma preserves strong multimodal understanding ability and compares favorably to popular large multimodal models that are trained on much larger datasets. We have made our model and code public for reproducibility at https://microsoft.github.io/Magma.

****

SimMotionEdit: Text-Based Human Motion Editing with Motion Similarity Prediction

Zhengyuan Li, Kai Cheng, Anindita Ghosh, Uttaran Bhattacharya, Liangyan Gui, Aniket Bera; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27827-27837

Text-based 3D human motion editing is a critical yet challenging task in computer vision and graphics. While training-free approaches have been explored, the recent release of the MotionFix dataset, which includes source-text-motion triplets, has opened new avenues for training, yielding promising results. However, existing methods struggle with precise control, often leading to misalignment between motion semantics and language instructions. In this paper, we introduce a related task --- motion similarity prediction --- and propose a multi-task training paradigm, where we train the model jointly on motion editing and motion similarity prediction to foster the learning of semantically meaningful representations. To complement this task, we design an advanced Diffusion-Transformer-based architecture that separately handles motion similarity prediction and motion editing. Extensive experiments demonstrate the state-of-the-art performance of our approach in both editing alignment and fidelity.

****

Object-aware Sound Source Localization via Audio-Visual Scene Understanding

Sung Jin Um, Dongjin Kim, Sangmin Lee, Jung Uk Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8342-8351

Audio-visual sound source localization task aims to spatially localize sound-making objects within visual scenes by integrating visual and audio cues. However, existing methods struggle with accurately localizing sound-making objects in complex scenes, particularly when visually similar silent objects coexist. This limitation arises primarily from their reliance on simple audio-visual correspondence, which does not capture fine-grained semantic differences between sound-making and silent objects. To address these challenges, we propose a novel sound source localization framework leveraging Multimodal Large Language Models (MLLMs) to generate detailed contextual information that explicitly distinguishes between sound-making foreground objects and silent background objects. To effectively integrate this detailed information, we introduce two novel loss functions: Object-aware Contrastive Alignment (OCA) loss and Object Region Isolation (ORI) loss. Extensive experimental results on MUSIC and VGGSound datasets demonstrate the effectiveness of our approach, significantly outperforming existing methods in both single-source and multi-source localization scenarios. Code and generated detailed contextual information are available at: https://github.com/VisualAIKHU/OA-SSL.

****
Volume Tells: Dual Cycle-Consistent Diffusion for 3D Fluorescence Microscopy Denoising and Super-Resolution
Zelin Li, Chenwei Wang, Zhaoke Huang, Yiming Ma, Cunming Zhao, Zhongying Zhao, Hong Yan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16091-16100
3D fluorescence microscopy is essential for understanding fundamental life processes through long-term live-cell imaging. However, due to inherent issues in imaging principles, it faces significant challenges including spatially varying noise and anisotropic resolution, where the axial resolution lags behind the lateral resolution up to 4.5 times. Meanwhile, laser power is kept low to maintain cell viability, leading to inaccessible low-noise and high-resolution paired ground truth (GT). To tackle these limitations, a dual Cycle-consistent Diffusion is proposed to effectively mine intra-volume imaging priors within 3D cell volumes in an unsupervised manner, i.e., Volume Tells (VTCD), achieving de-noising and super-resolution (SR) simultaneously. Specifically, a spatially iso-distributed denoiser is designed to exploit the noise distribution consistency between adjacent low-noise and high-noise regions within the 3D cell volume, suppressing the spatially varying noise. Then, in light of the structural consistency of the cell volume, a cross-plane global-propagation SR module propagates high-resolution details from the XY plane into adjacent regions in the XZ and YZ planes, progressively enhancing resolution across the entire 3D cell volume. Experimental results on 10 in vivo cellular dataset demonstrate high improvements in both denoising and super-resolution, with axial resolution enhanced from 430nm to 90nm
****
SerialGen: Personalized Image Generation by First Standardization Then Personalization
Cong Xie, Han Zou, Ruiqi Yu, Yan Zhang, Zhenpeng Zhan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2847-2856
In this work, we are interested in achieving both high text controllability and whole-body appearance consistency in the generation of personalized human characters. We propose a novel framework, named SerialGen, which is a serial generation method consisting of two stages: first, a standardization stage that standardizes reference images, and then a personalized generation stage based on the standardized reference. Furthermore, we introduce two modules aimed at enhancing the standardization process. Our experimental results validate the proposed framework's ability to produce personalized images that faithfully recover the reference image's whole-body appearance while accurately responding to a wide range of text prompts. Through thorough analysis, we highlight the critical contribution of the proposed serial generation method and standardization model, evidencing enhancements in appearance consistency between reference and output images and across serial outputs generated from diverse text prompts. The term "Serial" in this work carries a double meaning: it refers to the two-stage method and also underlines our ability to generate serial images with consistent appearance throughout.
****
From Head to Tail: Efficient Black-box Model Inversion Attack via Long-tailed Learning
Ziang Li, Hongguang Zhang, Juan Wang, Meihui Chen, Hongxin Hu, Wenzhe Yi, Xiaoyang Xu, Mengda Yang, Chenjun Ma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29288-29298
Model Inversion Attacks (MIAs) aim to reconstruct private training data from models, leading to privacy leakage, particularly in facial recognition systems. Although many studies have enhanced the effectiveness of white-box MIAs, less attention has been paid to improving efficiency and utility under limited attacker capabilities. Existing black-box MIAs necessitate an impractical number of queries, incurring significant overhead. Therefore, we analyze the limitations of existing MIAs and introduce Surrogate Model-based Inversion with Long-tailed Enhancement (SMILE), a high-resolution oriented and query-efficient MIA for the black-box setting. We begin by analyzing the initialization of MIAs from a data distribu

tion perspective and propose a long-tailed surrogate training method to obtain h
igh-quality initial points. We then enhance the attack's effectiveness by employ
ing the gradient-free black-box optimization algorithm selected by NGOpt. Our ex
periments show that SMILE outperforms existing state-of-the-art black-box MIAs w
hile requiring only about 5% of the query overhead.
****

Augmented Deep Contexts for Spatially Embedded Video Coding
Yifan Bian, Chuanbo Tang, Li Li, Dong Liu; Proceedings of the Computer Vision an
d Pattern Recognition Conference (CVPR), 2025, pp. 2094-2104
Most Neural Video Codecs (NVCs) only employ temporal references to generate temp
oral-only contexts and latent prior. These temporal-only NVCs fail to handle lar
ge motions or emerging objects due to limited contexts and misaligned latent pri
or. To relieve the limitations, we propose a Spatially Embedded Video Codec (SEV
C), in which the low-resolution video is compressed for spatial references. Firs
tly, our SEVC leverages both spatial and temporal references to generate augment
ed motion vectors and hybrid spatial-temporal contexts. Secondly, to address the
 misalignment issue in latent prior and enrich the prior information, we introdu
ce a spatial-guided latent prior augmented by multiple temporal latent represent
ations. At last, we design a joint spatial-temporal optimization to learn qualit
y-adaptive bit allocation for spatial references, further boosting rate-distorti
on performance. Experimental results show that our SEVC effectively alleviates t
he limitations in handling large motions or emerging objects, and also reduces 1
1.9% more bitrate than the previous state-of-the-art NVC while providing an addi
tional low-resolution bitstream. Our code and model are available at https://git
hub.com/EsakaK/SEVC.
****

SIDA: Social Media Image Deepfake Detection, Localization and Explanation with L
arge Multimodal Model
Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan
 Wu, Xiaowei Huang, Guangliang Cheng; Proceedings of the Computer Vision and Pat
tern Recognition Conference (CVPR), 2025, pp. 28831-28841
The rapid advancement of generative models in creating highly realistic images p
oses substantial risks for misinformation dissemination. For instance, a synthet
ic image, when shared on social media, can mislead extensive audiences and erode
 trust in digital content, resulting in severe repercussions. Despite some progr
ess, academia has not yet created a large and diversified deepfake detection dat
aset for social media, nor has it devised an effective solution to address this
issue. In this paper, we introduce the Social media Image Detection dataSet (SID
-Set), which offers three key advantages: (1) extensive volume, featuring 300K A
I-generated/tampered and authentic images with comprehensive annotations, (2) br
oad diversity, encompassing fully synthetic and tampered images across various c
lasses, and (3) elevated realism, with images that are predominantly indistingui
shable from genuine ones through mere visual inspection. Furthermore, leveraging
 the exceptional capabilities of large multimodal models, we propose a new image
 deepfake detection, localization, and explanation framework, named SIDA (Social
 media Image Detection, localization, and explanation Assistant). SIDA not only
discerns the authenticity of images, but also delineates tampered regions throug
h mask prediction and provides textual explanations of the model's judgment crit
eria. Compared with state-of-the-art deepfake detection models on SID-Set and ot
her benchmarks, extensive experiments demonstrate that SIDA achieves superior pe
rformance among diversified settings. The code, model, and dataset will be relea
sed.
****

Matrix3D: Large Photogrammetry Model All-in-One
Yuanxun Lu, Jingyang Zhang, Tian Fang, Jean-Daniel Nahmias, Yanghai Tsin, Long Q
uan, Xun Cao, Yao Yao, Shiwei Li; Proceedings of the Computer Vision and Pattern
 Recognition Conference (CVPR), 2025, pp. 11250-11263
We present Matrix3D, a unified model that performs several photogrammetry subtas
ks, including pose estimation, depth prediction, and novel view synthesis using
just the same model. Matrix3D utilizes a multi-modal diffusion transformer (DiT)

to integrate transformations across several modalities, such as images, camera parameters, and depth maps. The key to Matrix3D's large-scale multi-modal training lies in the incorporation of a mask learning strategy. This enables full-modality model training even with partially complete data, such as bi-modality data of image-pose and image-depth pairs, thus significantly increases the pool of available training data. Matrix3D demonstrates state-of-the-art performance in pose estimation and novel view synthesis tasks. Additionally, it offers fine-grained control through multi-round interactions, making it an innovative tool for 3D content creation. Project page: https://nju-3dv.github.io/projects/matrix3d.

****

Object-Centric Prompt-Driven Vision-Language-Action Model for Robotic Manipulation

Xiaoqi Li, Jingyun Xu, Mingxu Zhang, Jiaming Liu, Yan Shen, Iaroslav Ponomarenko, Jiahui Xu, Liang Heng, Siyuan Huang, Shanghang Zhang, Hao Dong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27638-27648

In robotic manipulation, task goals can be conveyed through various modalities, such as language, goal images, and goal videos. However, natural language can be ambiguous, while images or videos may offer overly detailed specifications. To address these challenges, we propose a novel approach using comprehensive multi-modal prompts that explicitly convey both low-level actions and high-level planning in a simple manner. Specifically, for each key-frame in the task sequence, our method allows for manual or automatic generation of simple and expressive 2D visual prompts overlaid on RGB images. These prompts represent the required task goals, such as the end-effector pose and the desired movement direction after contact. We develop a training strategy that enables the model to interpret these visual-language prompts and predict the corresponding contact poses and movement directions in SE(3) space. Furthermore, by sequentially executing all key-frame steps, the model can complete long-horizon tasks. This approach not only helps the model explicitly understand the task objectives but also enhances its robustness on unseen tasks by providing easily interpretable prompts. We evaluate our method in both simulated and real-world environments, demonstrating its robust manipulation capabilities.

****

Proximal Algorithm Unrolling: Flexible and Efficient Reconstruction Networks for Single-Pixel Imaging

Ping Wang, Lishun Wang, Gang Qu, Xiaodong Wang, Yulun Zhang, Xin Yuan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 411-421

Deep-unrolling and plug-and-play (PnP) approaches have become the de-facto standard solvers for single-pixel imaging (SPI) inverse problem. PnP approaches, a class of iterative algorithms where regularization is implicitly performed by an off-the-shelf deep denoiser, are flexible for varying compression ratios (CRs) but are limited in reconstruction accuracy and speed. Conversely, unrolling approaches, a class of multi-stage neural networks where a truncated iterative optimization process is transformed into an end-to-end trainable network, typically achieve better accuracy with faster inference but require fine-tuning or even retraining when CR changes. In this paper, we address the challenge of integrating the strengths of both classes of solvers. To this end, we design an efficient deep image restorer (DIR) for the unrolling of HQS (half quadratic splitting) and ADMM (alternating direction method of multipliers). More importantly, a general proximal trajectory (PT) loss function is proposed to train HQS/ADMM-unrolling networks such that learned DIR approximates the proximal operator of an ideal explicit restoration regularizer. Extensive experiments demonstrate that, the resulting proximal unrolling networks can not only flexibly handle varying CRs with a single model like PnP algorithms, but also outperform previous CR-specific unrolling networks in both reconstruction accuracy and speed. Source codes and models are available at https://github.com/pwangcs/ProxUnroll.

****

3DEnhancer: Consistent Multi-View Diffusion for 3D Enhancement

Yihang Luo, Shangchen Zhou, Yushi Lan, Xingang Pan, Chen Change Loy; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16430-16440

Despite advances in neural rendering, due to the scarcity of high-quality 3D datasets and the inherent limitations of multi-view diffusion models, view synthesis and 3D model generation are restricted to low resolutions with suboptimal multi-view consistency. In this study, we present a novel 3D enhancement pipeline, dubbed 3DEnhancer, which employs a multi-view latent diffusion model to enhance coarse 3D inputs while preserving multi-view consistency. Our method includes a pose-aware encoder and a diffusion-based denoiser to refine low-quality multi-view images, along with data augmentation and a multi-view attention module with epipolar aggregation to maintain consistent, high-quality 3D outputs across views. Unlike existing video-based approaches, our model supports seamless multi-view enhancement with improved coherence across diverse viewing angles. Extensive evaluations show that 3DEnhancer significantly outperforms existing methods, boosting both multi-view enhancement and per-instance 3D optimization tasks.

****

Investigating the Role of Weight Decay in Enhancing Nonconvex SGD

Tao Sun, Yuhao Huang, Li Shen, Kele Xu, Bao Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15287-15296

Weight decay is a widely used technique in training machine learning models, known to empirically enhance the generalization of Stochastic Gradient Descent (SGD). While intuitively weight decay allows SGD to train a regularized model rather than the original one, there is limited theoretical understanding of why SGD with weight decay (SGDW) yields results consistent with the unregularized model, or how weight decay improves generalization. This paper establishes a convergence theory for SGDW in the context of the unregularized model, under weaker assumptions than previous analyses of weight decay. Our theory demonstrates that weight decay does not accelerate the convergence of SGD. For generalization, we provide the first theoretical proof of weight decay's benefit in nonconvex optimization. Additionally, we extend our results to sign-based stochastic gradient algorithms, such as SignSGD. Numerical experiments on classical benchmarks validate our theoretical findings.

****

MarkushGrapher: Joint Visual and Textual Recognition of Markush Structures

Lucas Morin, Valery Weber, Ahmed Nassar, Gerhard Ingmar Meijer, Luc Van Gool, Yawei Li, Peter Staar; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14505-14515

The automated analysis of chemical literature holds promise to accelerate discovery in fields such as material science and drug development. In particular, search capabilities for chemical structures and Markush structures (chemical structure templates) within patent documents are valuable, e.g., for prior-art search. Advancements have been made in the automatic extraction of chemical structures from text and images, yet the Markush structures remain largely unexplored due to their complex multi-modal nature. In this work, we present MarkushGrapher, a multi-modal approach for recognizing Markush structures in documents. Our method jointly encodes text, image, and layout information through a Vision-Text-Layout encoder and an Optical Chemical Structure Recognition vision encoder. These representations are merged and used to auto-regressively generate a sequential graph representation of the Markush structure along with a table defining its variable groups. To overcome the lack of real-world training data, we propose a synthetic data generation pipeline that produces a wide range of realistic Markush structures. Additionally, we present M2S, the first annotated benchmark of real-world Markush structures, to advance research on this challenging task. Extensive experiments demonstrate that our approach outperforms state-of-the-art chemistry-specific and general-purpose vision-language models in most evaluation settings. Code, models, and datasets are available.

****

Depth Any Camera: Zero-Shot Metric Depth Estimation from Any Camera

Yuliang Guo, Sparsh Garg, S. Mahdi H. Miangoleh, Xinyu Huang, Liu Ren; Proceedin

gs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26996-27006
While recent depth foundation models exhibit strong zero-shot generalization, achieving accurate metric depth across diverse camera types--particularly those with large fields of view (FoV) such as fisheye and 360-degree cameras--remains a significant challenge. This paper presents Depth Any Camera (DAC), a powerful zero-shot metric depth estimation framework that extends a perspective-trained model to effectively handle cameras with varying FoVs. The framework is designed to ensure that all existing 3D data can be leveraged, regardless of the specific camera types used in new applications. Remarkably, DAC is trained exclusively on perspective images but generalizes seamlessly to fisheye and 360-degree cameras without the need for specialized training data. DAC employs Equi-Rectangular Projection (ERP) as a unified image representation, enabling consistent processing of images with diverse FoVs. Its key components include a pitch-aware Image-to-ERP conversion for efficient online augmentation in ERP space, a FoV alignment operation to support effective training across a wide range of FoVs, and multi-resolution data augmentation to address resolution disparities between training and testing. DAC achieves state-of-the-art zero-shot metric depth estimation, improving delta-1 accuracy by up to 50% on multiple fisheye and 360-degree datasets compared to prior metric depth foundation models, demonstrating robust generalization across camera types.
****

Image Quality Assessment: From Human to Machine Preference
Chunyi Li, Yuan Tian, Xiaoyue Ling, Zicheng Zhang, Haodong Duan, Haoning Wu, Zicheng Jia, Xiaohong Liu, Xiongkuo Min, Guo Lu, Weisi Lin, Guangtao Zhai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7570-7581
Image Quality Assessment (IQA) based on human subjective preferences has undergone extensive research in the past decades. However, with the development of communication protocols, the visual data consumption volume of machines has gradually surpassed that of humans. For machines, the preference depends on downstream tasks such as segmentation and detection, rather than visual appeal. Considering the huge gap between human and machine vision systems, this paper proposes the topic: Image Quality Assessment for Machine Vision for the first time. Specifically, we (1) defined the subjective preferences of machines, including downstream tasks, test models, and evaluation metrics; (2) established the Machine Preference Database (MPD), which contains 2.25M fine-grained annotations and 30k reference/distorted image pair instances; (3) verified the performance of mainstream IQA algorithms on MPD. Experiments show that current IQA metrics are human-centric and cannot accurately characterize machine preferences. We sincerely hope that MPD can promote the evolution of IQA from human to machine preferences. Project page is on: https://github.com/lcysyzxdxc/MPD.
****

ShotAdapter: Text-to-Multi-Shot Video Generation with Diffusion Models
Ozgur Kara, Krishna Kumar Singh, Feng Liu, Duygu Ceylan, James M. Rehg, Tobias Hinz; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28405-28415
Current diffusion-based text-to-video methods are limited to producing short video clips of a single shot and lack the capability to generate multi-shot videos with discrete transitions where the same character performs distinct activities across the same or different backgrounds. To address this limitation we propose a framework that includes a dataset collection pipeline and architectural extensions to video diffusion models to enable text-to-multi-shot video generation. Our approach enables generation of multi-shot videos as a single video with full attention across all frames of all shots, ensuring character and background consistency, and allows users to control the number, duration, and content of shots through shot-specific conditioning. This is achieved by incorporating a transition token into the text-to-video model to control at which frames a new shot begins and a local attention masking strategy which controls the transition token's effect and allows shot-specific prompting. To obtain training data we propose a n

ovel data collection pipeline to construct a multi-shot video dataset from existing single-shot video datasets. Extensive experiments demonstrate that fine-tuning a pre-trained text-to-video model for a few thousand iterations is enough for the model to subsequently be able to generate multi-shot videos with shot-specific control, outperforming the baselines. You can find more details in our webpage.
****

Context-Aware Multimodal Pretraining
Karsten Roth, Zeynep Akata, Dima Damen, Ivana Balazevic, Olivier J. Henaff; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4267-4279
Large-scale multimodal representation learning successfully optimizes for zero-shot transfer at test time. Yet the standard pretraining paradigm (contrastive learning on large amounts of image-text data) does not explicitly encourage representations to support few-shot adaptation. In this work, we propose a simple, but carefully designed extension to multimodal pretraining which enables representations to accommodate additional context. Using this objective, we show that vision-language models can be trained to exhibit significantly increased few-shot adaptation: across 21 downstream tasks, we find up to four-fold improvements in test-time sample efficiency, and average few-shot adaptation gains of over 5%, while retaining zero-shot generalization performance across model scales and training durations. In particular, equipped with simple, training-free, metric-based adaptation mechanisms, our representations surpass significantly more complex optimization-based adaptation schemes.
****

Sound Bridge: Associating Egocentric and Exocentric Videos via Audio Cues
Sihong Huang, Jiaxin Wu, Xiaoyong Wei, Yi Cai, Dongmei Jiang, Yaowei Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28942-28951
Understanding human behavior and the environmental information in the egocentric video is very challenging due to the invisibility of some actions (e.g., laughing and sneezing) and the local nature of the first-person view. Leveraging the corresponding exocentric video to provide global context has shown promising results. However, existing visual-to-visual and visual-to-textual Ego-Exo video alignment methods struggle with the problem that there could be non-visual overlap for the same activity. To address this, we propose using sound as a bridge, as audio is often consistent across Ego-Exo videos. However, direct audio-to-audio alignment lacks context. Thus, we introduce two context-aware sound modules: one aligns audio with vision via a visual-audio cross-attention module, and another aligns text with sound closed caption generated by LLM. Experimental results on two Ego-Exo video association benchmarks show that either of the two proposed modules manages to improve the state-of-the-art methods. Moreover, the proposed sound-aware egocentric or exocentric representation boosts the performance of downstream tasks, such as action recognition of exocentric videos and scene recognition of egocentric videos. The code and models can be accessed at https://github.com/shhuangcoder/SoundBridge.
****

Detecting Backdoor Attacks in Federated Learning via Direction Alignment Inspection
Jiahao Xu, Zikai Zhang, Rui Hu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20654-20664
The distributed nature of training makes Federated Learning (FL) vulnerable to backdoor attacks, where malicious model updates aim to compromise the global model's performance on specific tasks. Existing defense methods show limited efficacy as they overlook the inconsistency between benign and malicious model updates regarding both general and fine-grained directions. To fill this gap, we introduce AlignIns, a novel defense method designed to safeguard FL systems against backdoor attacks. AlignIns looks into the direction of each model update through a direction alignment inspection process. Specifically, it examines the alignment of model updates with the overall update direction and analyzes the distribution

of the signs of their significant parameters, comparing them with the principle sign across all model updates. Model updates that exhibit an unusual degree of alignment are considered malicious and thus be filtered out. We provide the theoretical analysis of the robustness of AlignIns and its propagation error in FL. Our empirical results on both independent and identically distributed (IID) and non-IID datasets demonstrate that AlignIns achieves higher robustness compared to the state-of-the-art defense methods. Code is available at https://anonymous.4open.science/r/AlignIns.
****

OmniDocBench: Benchmarking Diverse PDF Document Parsing with Comprehensive Annotations

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, Jin Shi, Fan Wu, Pei Chu, Minghao Liu, Zhenxiang Li, Chao Xu, Bo Zhang, Botian Shi, Zhongying Tu, Conghui He; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24838-24848

Document content extraction is a critical task in computer vision, underpinning the data needs of large language models (LLMs) and retrieval-augmented generation (RAG) systems. Despite recent progress, current document parsing methods have not been fairly and comprehensively evaluated due to the narrow coverage of document types and the simplified, unrealistic evaluation procedures in existing benchmarks. To address these gaps, we introduce OmniDocBench, a novel benchmark featuring high-quality annotations across nine document sources, including academic papers, textbooks, and more challenging cases such as handwritten notes and densely typeset newspapers. OmniDocBench supports flexible, multi-level evaluations--ranging from an end-to-end assessment to the task-specific and attribute-based analysis--using 19 layout categories and 15 attribute labels. We conduct a thorough evaluation of both pipeline-based methods and end-to-end vision-language models, revealing their strengths and weaknesses across different document types. OmniDocBench sets a new standard for the fair, diverse, and fine-grained evaluation in document parsing. Dataset and code are available at https://github.com/opendatalab/OmniDocBench.
****

LayoutVLM: Differentiable Optimization of 3D Layout via Vision-Language Models

Fan-Yun Sun, Weiyu Liu, Siyi Gu, Dylan Lim, Goutam Bhat, Federico Tombari, Manling Li, Nick Haber, Jiajun Wu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29469-29478

Spatial reasoning is a fundamental aspect of human cognition, enabling intuitive understanding and manipulation of objects in three-dimensional space. While foundation models demonstrate remarkable performance on some benchmarks, they still struggle with 3D reasoning tasks like arranging objects in space according to open-ended language instructions, particularly in dense and physically constrained environments. We introduce LayoutVLM, a framework and scene layout representation that exploits the semantic knowledge of Vision-Language Models (VLMs) and supports differentiable optimization to ensure physical plausibility. LayoutVLM employs VLMs to generate two mutually reinforcing representations from visually marked images, and a self-consistent decoding process to improve VLMs spatial planning. Our experiments show that LayoutVLM addresses the limitations of existing LLM and constraint-based approaches, producing physically plausible 3D layouts better aligned with the semantic intent of input language instructions. We also demonstrate that fine-tuning VLMs with the proposed scene layout representation extracted from existing scene datasets can improve their reasoning performance.
****

Point Clouds Meets Physics: Dynamic Acoustic Field Fitting Network for Point Cloud Understanding

Changshuo Wang, Shuting He, Xiang Fang, Jiawei Han, Zhonghang Liu, Xin Ning, Weijun Li, Prayag Tiwari; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22182-22192

While existing pre-training-based methods have enhanced point cloud model performance, they have not fundamentally resolved the challenge of local structure rep

resentation in point clouds. The limited representational capacity of pure point cloud models continues to constrain the potential of cross-modal fusion methods and performance across various tasks. To address this challenge, we propose a Dynamic Acoustic Field Fitting Network (DAF-Net), inspired by physical acoustic principles. Specifically, we represent local point clouds as acoustic fields and introduce a novel Acoustic Field Convolution (AF-Conv), which treats local aggregation as an acoustic energy field modeling problem and captures fine-grained local shape awareness by dividing the local area into near field and far field. Furthermore, drawing inspiration from multi-frequency wave phenomena and dynamic convolution, we develop the Dynamic Acoustic Field Convolution (DAF-Conv) based on AF-Conv. DAF-Conv dynamically generates multiple weights based on local geometric priors, effectively enhancing adaptability to diverse geometric features. Additionally, we design a Global Shape-Aware (GSA) layer incorporating EdgeConv and multi-head attention mechanisms, which combines with DAF-Conv to form the DAF Block. These blocks are then stacked to create a hierarchical DAFNet architecture. Extensive experiments demonstrate that DAFNet significantly outperforms existing methods across multiple tasks.

****

Faster Parameter-Efficient Tuning with Token Redundancy Reduction

Kwonyoung Kim, Jungin Park, Jin Kim, Hyeongjun Kwon, Kwanghoon Sohn; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30189-30198

Parameter-efficient tuning (PET) aims to transfer pre-trained foundation models to downstream tasks by learning a small number of parameters. Compared to traditional fine-tuning, which updates the entire model, PET significantly reduces storage and transfer costs for each task regardless of exponentially increasing pre-trained model capacity. However, most PET methods inherit the inference latency of their large backbone models and often introduce additional computational overhead due to additional modules (e.g. adapters), limiting their practicality for compute-intensive applications. In this paper, we propose Faster Parameter-Efficient Tuning (FPET), a novel approach that enhances inference speed and training efficiency while maintaining high storage efficiency. Specifically, we introduce a plug-and-play token redundancy reduction module delicately designed for PET. This module refines tokens from the self-attention layer using an adapter to learn the accurate similarity between tokens and cuts off the tokens through a fully-differentiable token merging strategy, which uses a straight-through estimator for optimal token reduction. Experimental results prove that our FPET achieves faster inference and higher memory efficiency than the pre-trained backbone while keeping competitive performance on par with state-of-the-art PET methods.

****

BlockDance: Reuse Structurally Similar Spatio-Temporal Features to Accelerate Diffusion Transformers

Hui Zhang, Tingwei Gao, Jie Shao, Zuxuan Wu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12891-12900

Diffusion models have demonstrated impressive generation capabilities, particularly with recent advancements leveraging transformer architectures to improve both visual and artistic quality. However, Diffusion Transformers (DiTs) continue to encounter challenges related to low inference speed, primarily due to the iterative denoising process. To address this issue, we propose BlockDance, a training-free approach that explores feature similarities at adjacent time steps to accelerate DiTs. Unlike previous feature-reuse methods that lack tailored reuse strategies for features at different scales, BlockDance prioritizes the identification of the most structurally similar features, referred to as Structurally Similar Spatio-Temporal (STSS) features. These features are primarily located within the structure-focused blocks of the transformer during the later stages of denoising. BlockDance caches and reuses these highly similar features to mitigate redundant computation, thereby accelerating DiTs while maximizing consistency with the generated results of the original model. Furthermore, considering the diversity of generated content and the varying distributions of redundant features, we introduce BlockDance-Ada, a lightweight decision-making network tailored for in

stance-specific acceleration. BlockDance-Ada dynamically allocates resources and provides superior content quality. Both BlockDance and BlockDance-Ada have proven effective across various generation tasks and models, achieving accelerations between 25% and 50% while maintaining generation quality.

****

Panorama Generation From NFoV Image Done Right

Dian Zheng, Cheng Zhang, Xiao-Ming Wu, Cao Li, Chengfei Lv, Jian-Fang Hu, Wei-Shi Zheng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21610-21619

Generating 360-degree panoramas from narrow field of view (NFoV) image is a promising computer vision task for Virtual Reality (VR) applications. Existing methods mostly assess the generated panoramas with InceptionNet or CLIP based metrics, which tend to perceive the image quality and is not suitable for evaluating the distortion. In this work, we first propose a distortion-specific CLIP, named Distort-CLIP to accurately evaluate the panorama distortion and discover the "visual cheating" phenomenon in previous works (i.e., tending to improve the visual results by sacrificing distortion accuracy). This phenomenon arises because prior methods employ a single network to learn the distinct panorama distortion and content completion at once, which leads the model to prioritize optimizing the latter. To address the phenomenon, we propose PanoDecouple, a decoupled diffusion model framework, which decouples the panorama generation into distortion guidance and content completion, aiming to generate panoramas with both accurate distortion and visual appeal. Specifically, we design a DistortNet for distortion guidance by imposing panorama-specific distortion prior and a modified condition registration mechanism; and a ContentNet for content completion by imposing perspective image information. Additionally, a distortion correction loss function with Distort-CLIP is introduced to constrain the distortion explicitly. The extensive experiments validate that PanoDecouple surpasses existing methods both in distortion and visual metrics.

****

Mamba-Adaptor: State Space Model Adaptor for Visual Recognition

Fei Xie, Jiahao Nie, Yujin Tang, Wenkang Zhang, Hongshen Zhao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20124-20134

Recent State Space Models (SSM), especially Mamba, have demonstrated impressive performance in visual modeling and possess superior model efficiency. However, the application of Mamba to visual tasks suffers inferior performance due to three main constraints existing in the sequential model: 1) Casual computing is incapable of accessing global context; 2) Long-range forgetting occurs when computing the current hidden states; 3) Weak spatial structural modeling due to the transformed sequential input. To address these issues, we investigate a simple yet powerful vision task adapter for Mamba models, which consists of two functional modules: Adaptor-T and Adapator-S. When solving the hidden states for SSM, we apply a casual prediction module Adaptor-T to select a set of learnable locations as memory augmentation feature states to ease long-range forgetting issues. Moreover, we leverage Adapator-S, composed of multi-scale dilated convolutional kernels, to enhance the spatial modeling and introduce the image inductive bias into the feature output. Both modules can enlarge the context modeling in casual computing, as the output is enhanced by the inaccessible features. We explore three usages of Mamba-Adaptor: A general visual backbone for various vision tasks; A booster module to raise the performance of pretrained backbones; A highly efficient fine-tuning module that adapts the base model for transfer learning tasks. Extensive experiments verify the effectiveness of Mamba-Adapter in three settings. Notably, our Mamba-Adapter achieves state-of-the-art performance on the ImageNet and COCO benchmarks.

****

Robust Message Embedding via Attention Flow-Based Steganography

Huayuan Ye, Shenzhuo Zhang, Shiqi Jiang, Jing Liao, Shuhang Gu, Dejun Zheng, Changbo Wang, Chenhui Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12840-12849

Image steganography can hide information in a host image and obtain a stego imag
e that is perceptually indistinguishable from the original one. This technique h
as tremendous potential in scenarios like copyright protection and information r
etrospection. Some previous studies have proposed to enhance the robustness of t
he methods against image disturbances to increase their applicability. However,
they generally cannot achieve a satisfying balance between the steganography qua
lity and robustness. Instead of image-in-image steganography, we focus on the is
sue of message-in-image embedding that is robust to various real-world image dis
tortions. This task aims to embed information into a natural image and the decod
ing result is required to be completely accurate, which increases the difficulty
 of data concealing and revealing. Inspired by the recent developments in transf
ormer-based vision models, we discover that the tokenized representation of imag
e is naturally suitable for steganography task. In this paper, we propose a nove
l message embedding framework, called Robust Message Steganography (RMSteg), whi
ch is competent to hide message via QR Code in a host image based on an normaliz
ing flow-based model. The stego image derived by our method has imperceptible ch
anges and the encoded message can be accurately restored even if the image is pr
inted out and photographed. To our best knowledge, this is the first work that i
ntegrates the advantages of transformer models into normalizing flow. The code i
s available at https://github.com/huayuan4396/RMSteg.
****

Task-driven Image Fusion with Learnable Fusion Loss
Haowen Bai, Jiangshe Zhang, Zixiang Zhao, Yichen Wu, Lilun Deng, Yukun Cui, Tao
Feng, Shuang Xu; Proceedings of the Computer Vision and Pattern Recognition Conf
erence (CVPR), 2025, pp. 7457-7468
Multi-modal image fusion aggregates information from multiple sensor sources, ac
hieving superior visual quality and perceptual features compared to single-sourc
e images, often improving downstream tasks. However, current fusion methods for
downstream tasks still use predefined fusion objectives that potentially mismatc
h the downstream tasks, limiting adaptive guidance and reducing model flexibilit
y. To address this, we propose Task-driven Image Fusion (TDFusion), a fusion fra
mework incorporating a learnable fusion loss guided by task loss. Specifically,
our fusion loss includes learnable parameters modeled by a neural network called
 the loss generation module. This module is supervised by the downstream task lo
ss in a meta-learning manner. The learning objective is to minimize the task los
s of fused images after optimizing the fusion module with the fusion loss. Itera
tive updates between the fusion module and the loss module ensure that the fusio
n network evolves toward minimizing task loss, guiding the fusion process toward
 the task objectives. TDFusion's training relies entirely on the downstream task
 loss, making it adaptable to any specific task. It can be applied to any archit
ecture of fusion and task networks. Experiments demonstrate TDFusion's performan
ce through fusion experiments conducted on four different datasets, in addition
to evaluations on semantic segmentation and object detection tasks. The code is
available at https://github.com/HaowenBai/TDFusion.
****

Compositional Targeted Multi-Label Universal Perturbations
Hassan Mahmood, Ehsan Elhamifar; Proceedings of the Computer Vision and Pattern
Recognition Conference (CVPR), 2025, pp. 20580-20591
Generating targeted universal perturbations for multi-label recognition is a com
binatorially hard problem that requires exponential time and space complexity. T
o address the problem, we propose a compositional framework. We show that a simp
le independence assumption on label-wise universal perturbations naturally leads
 to an efficient optimization that requires learning affine convex cones spanned
 by label-wise universal perturbations, significantly reducing the problem compl
exity to linear time and space. During inference, the framework allows generatin
g universal perturbations for novel combinations of classes in constant time. We
 demonstrate the scalability of our method on large datasets and target sizes, e
valuating its performance on NUS-WIDE, MS-COCO, and OpenImages using state-of-th
e-art multi-label recognition models. Our results show that our approach outperf
orms baselines and achieves results comparable to methods with exponential compl

exity.
****
PatchGuard: Adversarially Robust Anomaly Detection and Localization through Visi on Transformers and Pseudo Anomalies
Mojtaba Nafez, Amirhossein Koochakian, Arad Maleki, Jafar Habibi, Mohammad Hosse in Rohban; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20383-20394
Anomaly Detection (AD) and Anomaly Localization (AL) are crucial in fields that demand high reliability, such as medical imaging and industrial monitoring. Howe ver, current AD and AL approaches are often susceptible to adversarial attacks d ue to limitations in training data, which typically include only normal, unlabel ed samples. This study introduces PatchGuard, an adversarially robust AD and AL method that incorporates pseudo anomalies with localization masks within a Visio n Transformer (ViT)-based architecture to address these vulnerabilities.We begin by examining the essential properties of pseudo anomalies, and follow it by pro viding theoretical insights into the attention mechanisms required to enhance th e adversarial robustness of AD and AL systems. We then present our approach, whi ch leverages Foreground-Aware Pseudo-Anomalies to overcome the deficiencies of p revious anomaly-aware methods. Our method incorporates these crafted pseudo-anom aly samples into a ViT-based framework, with adversarial training guided by a no vel loss function designed to improve model robustness, as supported by our theo retical analysis.Experimental results on well-established industrial and medical datasets demonstrate that PatchGuard significantly outperforms previous methods in adversarial settings, achieving performance gains of 53.2% in AD and 68.5% i n AL, while also maintaining competitive accuracy in non-adversarial settings.
****
Sparse Point Cloud Patches Rendering via Splitting 2D Gaussians
Changfeng Ma, Ran Bi, Jie Guo, Chongjun Wang, Yanwen Guo; Proceedings of the Com puter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27285-27294
Current learning-based methods predict NeRF or 3D Gaussians from point clouds to achieve photo-realistic rendering but still depend on categorical priors, dense point clouds, or additional refinements. Hence, we introduce a novel point clou d rendering method by predicting 2D Gaussians from point clouds. Our method inco rporates two identical modules with an entire-patch architecture enabling the ne twork to be generalized to multiple datasets. The module normalizes and initiali zes the Gaussians utilizing the point cloud information including normals, color s and distances. Then, splitting decoders are employed to refine the initial Gau ssians by duplicating them and predicting more accurate results, making our meth odology effectively accommodate sparse point clouds as well. Once trained, our a pproach exhibits direct generalization to point clouds across different categori es. The predicted Gaussians are employed directly for rendering without addition al refinement on the rendered images, retaining the benefits of 2D Gaussians. We conduct extensive experiments on various datasets, and the results demonstrate the superiority and generalization of our method, which achieves SOTA performanc e.
****
Distilling Monocular Foundation Model for Fine-grained Depth Completion
Yingping Liang, Yutao Hu, Wenqi Shao, Ying Fu; Proceedings of the Computer Visio n and Pattern Recognition Conference (CVPR), 2025, pp. 22254-22265
Depth completion involves predicting dense depth maps from sparse LiDAR inputs, a critical task for applications such as autonomous driving and robotics. Howeve r, sparse depth annotations from sensors limit the availability of dense supervi sion, which is necessary for learning detailed geometric features. To overcome t his limitation, we propose a two-stage knowledge distillation framework that lev erages powerful monocular foundation models to provide dense supervision for dep th completion. In the first stage, we introduce a pre-training strategy that gen erates diverse training data from natural images to distill geometric knowledge to depth completion. Specifically, we simulate LiDAR scans by utilizing monocula r depth and mesh reconstruction, thereby creating training data without requirin g ground-truth depth. Nonetheless, monocular depth estimation suffers from inher

ent scale ambiguity in real-world settings. To address this, in the second stage , we employ a scale- and shift-invariant loss (SSI Loss) to learn real-world sca les when fine-tuning on real-world datasets. Our two-stage distillation framewor k enables depth completion models to harness the strengths of monocular foundati on models. Experimental results show that models trained with our two-stage dist illation framework achieve top-ranked performance on the KITTI benchmark, demons trating improvements in both quantitative and qualitative metrics.
****

LamRA: Large Multimodal Model as Your Advanced Retrieval Assistant
Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfe ng Wang, Weidi Xie; Proceedings of the Computer Vision and Pattern Recognition C onference (CVPR), 2025, pp. 4015-4025
With the rapid advancement of multimodal information retrieval, increasingly com plex retrieval tasks have emerged. Existing methods predominately rely on task-s pecific fine-tuning of vision-language models, often those trained with image-te xt contrastive learning. In this paper, we explore the possibility of re-purposi ng generative Large Multimodal Models (LMMs) for retrieval. This approach enable s unifying all retrieval tasks under the same formulation and, more importantly, allows for extrapolation towards unseen retrieval tasks without additional trai ning. Our contributions can be summarised in the following aspects: (i) We intro duce LamRA, a versatile framework designed to empower LMMs with sophisticated re trieval and reranking capabilities. (ii) For retrieval, we adopt a two-stage tra ining strategy comprising language-only pre-training and multimodal instruction tuning to progressively enhance LMM's retrieval performance. (iii) For reranking , we employ joint training of both pointwise and listwise reranking, offering tw o distinct ways to further boost the retrieval performance. (iv) Extensive exper iments underscore the efficacy of our method in handling more than ten retrieval tasks, demonstrating robust performance in both supervised and zero-shot settin gs, including scenarios involving previously unseen retrieval tasks.
****

AniGrad: Anisotropic Gradient-Adaptive Sampling for 3D Reconstruction From Monoc ular Video
Noah Stier, Alex Rich, Pradeep Sen, Tobias Höllerer; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21814-21823
Recent image-based 3D reconstruction methods have achieved excellent quality for indoor scenes using 3D convolutional neural networks. However, they rely on a h igh-resolution grid in order to achieve detailed output surfaces, which is quite costly in terms of compute time, and it results in large mesh sizes that are mo re expensive to store, transmit, and render. In this paper we propose a new solu tion to this problem, using adaptive sampling. By re-formulating the final layer s of the network, we are able to analytically bound the local surface complexity , and set the local sample rate accordingly. Our method, AniGrad, achieves an or der of magnitude reduction in both surface extraction latency and mesh size, whi le preserving mesh accuracy and detail.
****

Neural Video Compression with Context Modulation
Chuanbo Tang, Zhuoyuan Li, Yifan Bian, Li Li, Dong Liu; Proceedings of the Compu ter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12553-12563
Efficient video coding is highly dependent on exploiting the temporal redundancy , which is usually achieved by extracting and leveraging the temporal context in the emerging conditional coding-based neural video codec (NVC). Although the la test NVC has achieved remarkable progress in improving the compression performan ce, the inherent temporal context propagation mechanism lacks the ability to suf ficiently leverage the reference information, limiting further improvement. In t his paper, we address the limitation by modulating the temporal context with the reference frame in two steps. Specifically, we first propose the flow orientati on to mine the inter-correlation between the reference frame and prediction fram e for generating the additional oriented temporal context. Moreover, we introduc e the context compensation to leverage the oriented context to modulate the prop agated temporal context generated from the propagated reference feature. Through

the synergy mechanism and decoupling loss supervision, the irrelevant propagated information can be effectively eliminated to ensure better context modeling. Experimental results demonstrate that our codec achieves on average 22.7% bitrate reduction over the advanced traditional video codec H.266/VVC, and offers an average 10.1% bitrate saving over the previous state-of-the-art NVC DCVC-FM. The code is available at https://github.com/Austin4USTC/DCMVC.
****

Less Attention is More: Prompt Transformer for Generalized Category Discovery
Wei Zhang, Baopeng Zhang, Zhu Teng, Wenxin Luo, Junnan Zou, Jianping Fan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30322-30331
Generalized Category Discovery (GCD) typically relies on the pre-trained Vision Transformer (ViT) to extract features from a global receptive field, followed by contrastive learning to simultaneously classify unlabeled known classes and unknown classes without priors. Owing to the deficiency in the modeling capacity for inner-patch local information within ViT, current methods primarily focus on discriminative features at global level. This results in a model with more yet scattered attention, where neither excessive nor insufficient focus can grasp subtle differences to classify fine-grained unknown and known categories. To address this issue, we propose the AptGCD to deliver apt attention for GCD. It mimics the human brain how leveraging visual perception to refine local attention and comprehend global context by proposing a Meta Visual Prompt (MVP) and Prompt Transformer (PT). MVP is introduced into GCD for the first time, refining channel-level attention, while adaptively self-learning unique inner-patch features as prompts to achieve local visual modeling for our prompt transformer. Yet, relying solely on detailed features can lead to skewed judgments. Hence, PT harmonizes local and global representations, guiding the model's interpretation of features through broader contexts, thereby capturing more useful details with less attention. Extensive experiments on seven datasets demonstrate that AptGCD outperforms current methods, it achieves an average proportional 'New' accuracy improvement of approximately 9.2% over SOTA method on the all four fine-grained datasets, establishing a new standard in the field. The code is available at https://github.com/wendy26zhang/AptGCD.
****

On-Device Self-Supervised Learning of Low-Latency Monocular Depth from Only Events
Jesse J. Hagenaars, Yilun Wu, Federico Paredes-Valles, Stein Stroobants, Guido C.H.E. de Croon; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17114-17123
Event cameras provide low-latency perception for only milliwatts of power. This makes them highly suitable for resource-restricted, agile robots such as small flying drones. Self-supervised learning based on contrast maximization holds great potential for event-based robot vision, as it foregoes the need for high-frequency ground truth and allows for online learning in the robot's operational environment. However, online, on-board learning raises the major challenge of achieving sufficient computational efficiency for real-time learning, while maintaining competitive visual perception performance. In this work, we improve the time and memory efficiency of the contrast maximization pipeline, making on-device learning of low-latency monocular depth possible. We demonstrate that online learning on board a small drone yields more accurate depth estimates and more successful obstacle avoidance behavior compared to only pre-training. Benchmarking experiments show that the proposed pipeline is not only efficient, but also achieves state-of-the-art depth estimation performance among self-supervised approaches. Our work taps into the unused potential of online, on-device robot learning, promising smaller reality gaps and better performance.
****

CoMM: A Coherent Interleaved Image-Text Dataset for Multimodal Understanding and Generation
Wei Chen, Lin Li, Yongqi Yang, Bin Wen, Fan Yang, Tingting Gao, Yu Wu, Long Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR),

2025, pp. 8073-8082
Interleaved image-text generation has emerged as a crucial multimodal task, aiming at creating sequences of interleaved visual and textual content given a query. Despite notable advancements in recent multimodal large language models (MLLMs), generating integrated image-text sequences that exhibit narrative coherence and entity and style consistency remains challenging due to poor training data quality. To address this gap, we introduce CoMM, a high-quality Coherent interleaved image-text MultiModal dataset designed to enhance the coherence, consistency, and alignment of generated multimodal content. Initially, CoMM harnesses raw data from diverse sources, focusing on instructional content and visual storytelling, establishing a foundation for coherent and consistent content. To further refine the data quality, we devise a multi-perspective filter strategy that leverages advanced pre-trained models to ensure the development of sentences, consistency of inserted images, and semantic alignment between them. Various quality evaluation metrics are designed to prove the high quality of the filtered dataset. Meanwhile, extensive few-shot experiments on various downstream tasks demonstrate CoMM's effectiveness in significantly enhancing the in-context learning capabilities of MLLMs. Moreover, we propose four new tasks to evaluate MLLMs' interleaved generation abilities, supported by a comprehensive evaluation framework. We believe CoMM opens a new avenue for advanced MLLMs with superior multimodal in-context learning and understanding ability. The dataset and codes will be released.
****
MoGe: Unlocking Accurate Monocular Geometry Estimation for Open-Domain Images with Optimal Training Supervision
Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, Jiaolong Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5261-5271
We present MoGe, a powerful model for recovering 3D geometry from monocular open-domain images. Given a single image, our model directly predicts a 3D point map of the captured scene with an affine-invariant representation, which is agnostic to true global scale and shift. This new representation precludes ambiguous supervision in training and facilitates effective geometry learning. Furthermore, we propose a set of novel global and local geometry supervision techniques that empower the model to learn high-quality geometry. These include a robust, optimal, and efficient point cloud alignment solver for accurate global shape learning, and a multi-scale local geometry loss promoting precise local geometry supervision. We train our model on a large, mixed dataset and demonstrate its strong generalizability and high accuracy. In our comprehensive evaluation on diverse unseen datasets, our model significantly outperforms state-of-the-art methods across all tasks, including monocular estimation of 3D point map, depth map, and camera field of view.
****
Beyond Background Shift: Rethinking Instance Replay in Continual Semantic Segmentation
Hongmei Yin, Tingliang Feng, Fan Lyu, Fanhua Shang, Hongying Liu, Wei Feng, Liang Wan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9839-9848
In this work, we focus on continual semantic segmentation (CSS), where segmentation networks are required to continuously learn new classes without erasing knowledge of previously learned ones. Although storing images of old classes and directly incorporating them into the training of new models has proven effective in mitigating catastrophic forgetting in classification tasks, this strategy presents notable limitations in CSS. Specifically, the stored and new images with partial category annotations leads to confusion between unannotated categories and the background, complicating model fitting. To tackle this issue, this paper proposes a novel Enhanced Instance Replay (EIR) method, which not only preserves knowledge of old classes while simultaneously eliminating background confusion by instance storage of old classes, but also mitigates background shifts in the new images by integrating stored instances with new images. By effectively resolvin

g background shifts in both stored and new images, EIR alleviates catastrophic f orgetting in the CSS task, thereby enhancing the model's capacity for CSS. Exper imental results validate the efficacy of our approach, which significantly outpe rforms state-of-the-art CSS methods.
****

ScaleLSD: Scalable Deep Line Segment Detection Streamlined
Zeran Ke, Bin Tan, Xianwei Zheng, Yujun Shen, Tianfu Wu, Nan Xue; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6327-6 336
This paper studies the problem of Line Segment Detection (LSD) for the character ization of line geometry in images, with the aim of learning a domain-agnostic r obust LSD model that works well for any natural images. With the focus of scalab le self-supervised learning of LSD, we revisit and streamline the fundamental de signs of (deep and non-deep) LSD approaches to have a high-performing and effici ent LSD learner, dubbed as ScaleLSD, for the curation of line geometry at scale from over 10M unlabeled real-world images. Our ScaleLSD works very well to detec t much more number of line segments from any natural images even than the pionee red non-deep LSD approach, having a more complete and accurate geometric charact erization of images using line segments. Experimentally, our proposed ScaleLSD i s comprehensively testified under the zero-shot protocol in detection performanc e, single-view 3D geometry estimation, two-view line segment matching, and multi view 3D line mapping, all with excellent perfor- mance obtained. Based on the th orough evaluation, our ScaleLSD is observed to be the first deep approach that o utperforms the pioneered non-deep LSD in all aspects we have tested, significant ly expanding and reinforcing the versatility of the line geometry of images.
****

AToM: Aligning Text-to-Motion Model at Event-Level with GPT-4Vision Reward
Haonan Han, Xiangzuo Wu, Huan Liao, Zunnan Xu, Zhongyuan Hu, Ronghui Li, Yachao Zhang, Xiu Li; Proceedings of the Computer Vision and Pattern Recognition Confer ence (CVPR), 2025, pp. 22746-22755
Recently, text-to-motion models open new possibilities for creating realistic hu man motion with greater efficiency and flexibility. However, aligning motion gen eration with event-level textual descriptions presents unique challenges due to the complex, nuanced relationship between textual prompts and desired motion out comes. To address this issue, we introduce AToM, a framework that enhances the a lignment between generated motion and text prompts by leveraging reward from GPT -4Vision. AToM comprises three main stages: Firstly, we construct a dataset Moti onPrefer that pairs three types of event-level textual prompts with generated mo tions, which cover the integrity, temporal relationship and the frequency of mot ion. Secondly, we design a paradigm that utilizes GPT-4Vision for detailed motio n annotation, including visual data formatting, task-specific instructions and s coring rules for each sub-task. Finally, we fine-tune an existing text-to-motion model using reinforcement learning guided by this paradigm. Experimental result s demonstrate that AToM significantly improves the event-level alignment quality of text-to-motion generation.
****

Revisiting MAE Pre-training for 3D Medical Image Segmentation
Tassilo Wald, Constantin Ulrich, Stanislav Lukyanenko, Andrei Goncharov, Alberto Paderno, Maximilian Miller, Leander Maerkisch, Paul Jaeger, Klaus Maier-Hein; P roceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 202 5, pp. 5186-5196
Self-Supervised Learning (SSL) presents an exciting opportunity to unlock the po tential of vast, untapped clinical datasets, for various downstream applications that suffer from the scarcity of labeled data. While SSL has revolutionized fie lds like natural language processing and computer vision, its adoption in 3D med ical image computing has been limited by three key pitfalls: Small pre-training dataset sizes, architectures inadequate for 3D medical image analysis, and insuf ficient evaluation practices. In this paper, we address these issues by i) lever aging a large-scale dataset of 39k 3D brain MRI volumes and ii) using a Residual Encoder U-Net architecture within the state-of-the-art nnU-Net framework. iii)

A robust development framework, incorporating 5 development and 8 testing brain MRI segmentation datasets, allowed performance-driven design decisions to optimize the simple concept of Masked Auto En- coders (MAEs) for 3D CNNs. The resulting model not only surpasses previous SSL methods but also outperforms the strong nnU-Net baseline by an average of approximately 3 Dice points setting a new state-of-the-art. Our code and models are made available here.
****

Learning with Noisy Triplet Correspondence for Composed Image Retrieval
Shuxian Li, Changhao He, Xiting Liu, Joey Tianyi Zhou, Xi Peng, Peng Hu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19628-19637
Composed Image Retrieval (CIR) enables editable image search by integrating a query pair--a reference image ref and a textual modification mod--to retrieve a target image tar that reflects the intended change. While existing CIR methods have shown promising performance using well-annotated triplets \langle ref, mod, tar \rangle, almost all of them implicitly assume these triplets are accurately associated with each other. In practice, however, this assumption is often violated due to the limited knowledge of annotators, inevitably leading to incorrect textual modifications and resulting in a practical yet less-touched problem: noisy triplet correspondence (NTC). To tackle this challenge, we propose a \underline T ask-oriented \underline M odification \underline E nhancement framework (TME) to learn robustly from noisy triplets, which comprises three key modules: Robust Fusion Query (RFQ), Pseudo Text Enhancement (PTE), and Task-Oriented Prompt (TOP). Specifically, to mitigate the adverse impact of noise, RFQ employs a sample selection strategy to divide the training triplets into clean and noisy sets, thus enhancing the reliability of the training data for robust learning. To further leverage the noisy data instead of discarding it, PTE unifies the triplet noise as an adapter mismatch problem, thereby adjusting mod to align with ref and tar in the mismatched triplet. Finally, TOP replaces ref in the clean set with a trainable prompt, which is then concatenated with mod to form a query independent of the visual reference, aiming to mitigate visually irrelevant noise. Extensive experiments on two domain-specific datasets demonstrate the robustness and superiority of TME, particularly in noisy scenarios.
****

DoF-Gaussian: Controllable Depth-of-Field for 3D Gaussian Splatting
Liao Shen, Tianqi Liu, Huiqiang Sun, Jiaqi Li, Zhiguo Cao, Wei Li, Chen Change Loy; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26462-26471
Recent advances in 3D Gaussian Splatting (3D-GS) have shown remarkable success in representing 3D scenes and generating high-quality, novel views in real-time. However, 3D-GS and its variants assume that input images are captured based on pinhole imaging and are fully in focus. This assumption limits their applicability, as real-world images often feature shallow depth-of-field (DoF). In this paper, we introduce DoF-Gaussian, a controllable depth-of-field method for 3D-GS. We develop a lens-based imaging model based on geometric optics principles to control DoF effects. To ensure accurate scene geometry, we incorporate depth priors adjusted per scene, and we apply defocus-to-focus adaptation to minimize the gap in the circle of confusion. We also introduce a synthetic dataset to assess refocusing capabilities and the model's ability to learn precise lens parameters. Our framework is customizable and supports various interactive applications. Extensive experiments confirm the effectiveness of our method. Our project is available at https://dof-gaussian.github.io/.
****

Parallelized Autoregressive Visual Generation
Yuqing Wang, Shuhuai Ren, Zhijie Lin, Yujin Han, Haoyuan Guo, Zhenheng Yang, Difan Zou, Jiashi Feng, Xihui Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12955-12965
Autoregressive models have emerged as a powerful approach for visual generation but suffer from slow inference speed due to their sequential token-by-token prediction process. In this paper, we propose a simple yet effective approach for pa

rallelized autoregressive visual generation that improves generation efficiency while preserving the advantages of autoregressive modeling. Our key insight is that parallel generation depends on visual token dependencies: tokens with weak dependencies can be generated in parallel, while strongly dependent adjacent tokens are difficult to generate together, as their independent sampling may lead to inconsistencies. Based on this observation, we develop a parallel generation strategy that generates distant tokens with weak dependencies in parallel while maintaining sequential generation for strongly dependent local tokens. Our approach can be seamlessly integrated into standard autoregressive models without modifying the architecture or tokenizer. Experiments on ImageNet and UCF-101 demonstrate that our method achieves a 3.6x speedup with comparable quality and up to 9.5x speedup with minimal quality degradation across both image and video generation tasks. We hope this work will inspire future research in efficient visual generation and unified autoregressive modeling. Project page: https://yuqingwang1029.github.io/PAR-project.

****

CGMatch: A Different Perspective of Semi-supervised Learning

Bo Cheng, Jueqing Lu, Yuan Tian, Haifeng Zhao, Yi Chang, Lan Du; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15381-15391

Semi-supervised learning (SSL) has garnered significant attention due to its ability to leverage limited labeled data and a large amount of unlabeled data to improve model generalization performance. Recent approaches achieve impressive successes by combining ideas from both consistency regularization and pseudo-labeling. However, these methods tend to underperform in the more realistic situations with relatively scarce labeled data. We argue that this issue arises because existing methods rely solely on the model's confidence, making them challenging to accurately assess the model's state and identify unlabeled examples contributing to the training phase when supervision information is limited, especially during the early stages of model training. In this paper, we propose a novel SSL model called CGMatch, which, for the first time, incorporates a new metric known as Count-Gap (CG). We demonstrate that CG is effective in discovering unlabeled examples beneficial for model training. Along with confidence, a commonly used metric in SSL, we propose a fine-grained dynamic selection (FDS) strategy. This strategy dynamically divides the unlabeled dataset into three subsets with different characteristics: easy-to-learn set, ambiguous set, and hard-to-learn set. By selective filtering subsets, and applying corresponding regularization with selected subsets, we mitigate the negative impact of incorrect pseudo-labels on model optimization and generalization. Extensive experimental results on several common SSL benchmarks indicate the effectiveness of CGMatch especially when the labeled data are particularly limited. Source code is available at https://github.com/BoCheng-96/CGMatch.

****

ChatHuman: Chatting about 3D Humans with Tools

Jing Lin, Yao Feng, Weiyang Liu, Michael J. Black; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8150-8161

Numerous methods have been proposed to detect, estimate, and analyze properties of people in images, including the estimation of 3D pose, shape, contact, human-object interaction, emotion, and more. While widely applicable in vision and other areas, such methods require expert knowledge to select, use, and interpret the results. To address this, we introduce ChatHuman, a language-driven system that combines and integrates the skills of these specialized methods. ChatHuman functions as an assistant proficient in utilizing, analyzing, and interacting with tools specific to 3D human tasks, adeptly discussing and resolving related challenges. Built on a Large Language Model (LLM) framework, ChatHuman is trained to autonomously select, apply, and interpret a diverse set of tools in response to user inputs. Adapting LLMs to 3D human tasks presents challenges, including the need for domain-specific knowledge and the ability to interpret complex 3D outputs. The innovative features of ChatHuman include leveraging academic publications to instruct the LLM how to use the tools, employing a retrieval-augmented gen

eration model to create in-context learning examples for managing new tools, and effectively discriminating and integrating tool results to enhance tasks related to 3D humans by transforming specialized 3D outputs into comprehensible formats. Our experiments demonstrate that ChatHuman surpasses existing models in both tool selection accuracy and overall performance across various 3D human tasks, and it supports interactive chatting with users. ChatHuman represents a significant step toward consolidating diverse analytical methods into a unified, robust system for 3D human tasks.
****

FIction: 4D Future Interaction Prediction from Video
Kumar Ashutosh, Georgios Pavlakos, Kristen Grauman; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17613-17625
Anticipating how a person will interact with objects in an environment is essential for activity understanding, but existing methods are limited to the 2D space of video frames--capturing physically ungrounded predictions of "what" and ignoring the "where" and "how". We introduce FIction for 4D future interaction prediction from videos. Given an input video of a human activity, the goal is to predict which objects at what 3D locations the person will interact with in the next time period (e.g., cabinet, fridge), and how they will execute that interaction (e.g., poses for bending, reaching, pulling). Our novel model FIction fuses the past video observation of the person's actions and their environment to predict both the "where" and "how" of future interactions. Through comprehensive experiments on a variety of activities and real-world environments in Ego-Exo4D, we show that our proposed approach outperforms prior autoregressive and (lifted) 2D video models substantially, with more than 30% relative gains.
****

D^2iT: Dynamic Diffusion Transformer for Accurate Image Generation
Weinan Jia, Mengqi Huang, Nan Chen, Lei Zhang, Zhendong Mao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12860-12870
Diffusion models are widely recognized for their ability to generate high-fidelity images. Despite the excellent performance and scalability of the Diffusion Transformer (DiT) architecture, it applies fixed compression across different image regions during the diffusion process, disregarding the naturally varying information densities present in these regions. However, large compression leads to limited local realism, while small compression increases computational complexity and compromises global consistency, ultimately impacting the quality of generated images. To address these limitations, we propose dynamically compressing different image regions by recognizing the importance of different regions, and introduce a novel two-stage framework designed to enhance the effectiveness and efficiency of image generation: (1) Dynamic VAE (DVAE) at first stage employs a hierarchical encoder to encode different image regions at different downsampling rates, tailored to their specific information densities, thereby providing more accurate and natural latent codes for the diffusion process. (2) Dynamic Diffusion Transformer (D^2iT) at second stage generates images by predicting multi-grained noise, consisting of coarse-grained (less latent code in smooth regions) and fine-grained (more latent codes in detailed regions), through an novel combination of the Dynamic Grain Transformer and the Dynamic Content Transformer. The strategy of combining rough prediction of noise with detailed regions correction achieves a unification of global consistency and local realism. Comprehensive experiments on various generation tasks validate the effectiveness of our approach. Code will be released at https://github.com/jiawn-creator/Dynamic-DiT.
****

Scalable Autoregressive Monocular Depth Estimation
Jinhong Wang, Jian Liu, Dongqi Tang, Weiqiang Wang, Wentong Li, Danny Chen, Jintai Chen, Jian Wu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6262-6272
This paper proposes a new autoregressive model as an effective and scalable monocular depth estimator. Our idea is simple: We tackle the monocular depth estimation (MDE) task with an autoregressive prediction paradigm, based on two core designs. First, our depth autoregressive model (DAR) treats the depth map of differ

ent resolutions as a set of tokens, and conducts the low-to-high resolution auto regressive objective with a patch-wise casual mask. Second, our DAR recursively discretizes the entire depth range into more compact intervals, and attains the coarse-to-fine granularity autoregressive objective in an ordinal-regression man ner. By coupling these two autoregressive objectives, our DAR establishes new st ate-of-the-art (SOTA) on KITTI and NYU Depth v2 by clear margins. Further, our s calable approach allows us to scale the model up to 2.0B and achieve the best RM SE of 1.799 on the KITTI dataset (5% improvement) compared to 1.896 by the curre nt SOTA (Depth Anything). DAR further showcases zero-shot generalization ability on unseen datasets. These results suggest that DAR yields superior performance with an autoregressive prediction paradigm, providing a promising approach to eq uip modern autoregressive large models (e.g., GPT-4o) with depth estimation capa bilities. Project page: https://depth-ar.github.io/

****

Reconciling Stochastic and Deterministic Strategies for Zero-shot Image Restorat ion using Diffusion Model in Dual

Chong Wang, Lanqing Guo, Zixuan Fu, Siyuan Yang, Hao Cheng, Alex C. Kot, Bihan W en; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) , 2025, pp. 23207-23216

Plug-and-play (PnP) methods offer an iterative strategy for solving image restor ation (IR) problems in a zero-shot manner, using a learned discriminative denois er as the implicit prior. More recently, a sampling-based variant of this approa ch, which utilizes a pre-trained generative diffusion model, has gained great po pularity for solving IR problems through stochastic sampling. The IR results usi ng PnP with a pre-trained diffusion model demonstrate distinct advantages compar ed to those using discriminative denoisers, i.e., improved perceptual quality wh ile sacrificing the data fidelity. The unsatisfactory results are due to the lac k of integration of these strategies in the IR tasks. In this work, we propose a novel zero-shot IR scheme, dubbed Reconciling Diffusion Model in Dual (RDMD), w hich leverages only a single pre-trained diffusion model to construct two comple mentary regularizers. Specifically, the diffusion model in RDMD will iteratively perform deterministic denoising and stochastic sampling, aiming to achieve high -fidelity image restoration with appealing perceptual quality.RDMD also allows u sers to customize the distortion-perception tradeoff with a single hyperparamete r, enhancing the adaptability of the restoration process in different practical scenarios. Extensive experiments on several IR tasks demonstrate that our propos ed method could achieve superior results compared to existing approaches on both the FFHQ and ImageNet datasets.

****

Hierarchical Flow Diffusion for Efficient Frame Interpolation

Yang Hai, Guo Wang, Tan Su, Wenjie Jiang, Yinlin Hu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22943-22952

Most recent diffusion-based methods still show a large gap compared to non-diffu sion methods for video frame interpolation, in both accuracy and efficiency. Mos t of them formulate the problem as a denoising procedure in latent space directl y, which is less effective caused by the large latent space. We propose to model bilateral optical flow explicitly by hierarchical diffusion models, which has m uch smaller search space in the denoising procedure. Based on the flow diffusion model, we then use a flow-guided image synthesizer to produce the final result. We train the flow diffusion model and the image synthesizer end to end. Our met hod achieves state of the art in accuracy, and 10+ times faster than other diffu sion-based methods. The project page is at: https://hfd-interpolation.github.io.

****

Recurrence-Enhanced Vision-and-Language Transformers for Robust Multimodal Docum ent Retrieval

Davide Caffagni, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara; P roceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 202 5, pp. 9286-9295

Cross-modal retrieval is gaining increasing efficacy and interest from the resea rch community, thanks to large-scale training, novel architectural and learning

designs, and its application in LLMs and multimodal LLMs. In this paper, we move a step forward and design an approach that allows for multimodal queries - composed of both an image and a text - and can search within collections of multimodal documents, where images and text are interleaved. Our model, ReT, employs multi-level representations extracted from different layers of both visual and textual backbones, both at the query and document side. To allow for multi-level and cross-modal understanding and feature extraction, ReT employs a novel Transformer-based recurrent cell that integrates both textual and visual features at different layers, and leverages sigmoidal gates inspired by the classical design of LSTMs. Extensive experiments on M2KR and M-BEIR benchmarks show that ReT achieves state-of-the-art performance across diverse settings. Our source code and trained models are publicly available at: https://github.com/aimagelab/ReT.

****

Camouflage Anything: Learning to Hide using Controlled Out-painting and Representation Engineering

Biplab Das, Viswanath Gopalakrishnan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3603-3613

In this work, we introduce Camouflage Anything, a novel and robust approach to generate camouflaged datasets. To the best of our knowledge, we are the first to apply Controlled Out-painting and Representation Engineering for generating realistic camouflaged images with an objective to hide any segmented object coming from a generic or salient database. Our proposed method uses a novel control design to out-paint a given segmented object, with a camouflaged background. We also uncover the role of representation engineering in enhancing the quality of generated camouflage datasets. We address the limitations of existing metrics FID and KID in capturing the 'camouflage quality' ,by proposing a novel metric namely, CamOT. CamOT uses Optimal Transport between foreground & background Gaussian Mixture Models (GMM) of concerned camouflaged object to assign an image quality score. Furthermore, we conduct LoRA-based fine-tuning of the robust BiRefNet baseline with our generated camouflaged datasets, leading to notable improvements in camouflaged object segmentation accuracy. The experimental results showcase the efficacy and potential of Camouflage Anything, outperforming existing methods in camouflaged generation tasks.

****

Test-Time Fine-Tuning of Image Compression Models for Multi-Task Adaptability

Unki Park, Seongmoon Jeong, Youngchan Jang, Gyeong-Moon Park, Jong Hwan Ko; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4430-4440

The field of computer vision was initially inspired by the human visual system and has progressively expanded to include a broader range of machine vision applications. Consequently, image compressors should be designed to effectively accommodate not only human visual perception but also machine vision tasks, including closed-set scenarios that enable pre-training and open-set scenarios that involve previously unseen tasks at test time. Many recent studies effectively address both human visual perception and closed-set machine vision tasks simultaneously but struggle to handle open-set machine vision tasks. To address this issue, this paper proposes a fully instance-specific test time fine-tuning (TTFT) for adapting learned image compression (LIC) to both closed-set and open-set machine vision tasks effectively. With our method, a large-scale LIC model, originally trained for human perception, is adapted to the target task through TTFT using Singular Value Decomposition based Low Rank Adaptation (SVD-LoRA). During TTFT, the decoder adopts a modified learning scheme that focuses exclusively on training the singular values, which helps prevent excessive bitstream overhead. This enables fully instance-specific optimization for the target task, even for open-set tasks. Experimental results demonstrate that the proposed method effectively adapts the backbone compressor to diverse machine vision tasks, outperforming competing methods.

****

BASKET: A Large-Scale Video Dataset for Fine-Grained Skill Estimation

Yulu Pan, Ce Zhang, Gedas Bertasius; Proceedings of the Computer Vision and Patt

ern Recognition Conference (CVPR), 2025, pp. 28952-28962
We present BASKET, a large-scale basketball video dataset for fine-grained skill estimation. BASKET contains 4,477 hours of video capturing 32,232 basketball players from all over the world. Compared to prior skill estimation datasets, our dataset includes a massive number of skilled participants with unprecedented diversity in terms of gender, age, skill level, geographical location, etc. BASKET includes 20 fine-grained basketball skills, challenging modern video recognition models to capture the intricate nuances of player skill through in-depth video analysis. Given a long highlight video (8-10 minutes) of a particular player, the model needs to predict the skill level (e.g., excellent, good, average, fair, poor) for each of the 20 basketball skills. Our empirical analysis reveals that the current state-of-the-art video models struggle with this task, significantly lagging behind the human baseline. We believe that BASKET could be a useful resource for developing new video models with advanced long-range, fine-grained recognition capabilities. In addition, we hope that our dataset will be useful for domain-specific applications such as fair basketball scouting, personalized player development, and many others. Dataset and code are available at https://github.com/yulupan00/BASKET.
****

AniDoc: Animation Creation Made Easier
Yihao Meng, Hao Ouyang, Hanlin Wang, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Zhiheng Liu, Yujun Shen, Huamin Qu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18187-18197
The production of 2D animation follows an industry-standard workflow, encompassing four essential stages: character design, keyframe animation, in-betweening, and coloring. Our research focuses on reducing the labor costs in the above process by harnessing the potential of increasingly powerful generative AI. Using video diffusion models as the foundation, Anidoc emerges as a video line art colorization tool, which automatically converts sketch sequences into colored animations following the reference character specification. Our model exploits correspondence matching as an explicit guidance, yielding strong robustness to the variations (e.g., posture) between the reference character and each line art frame. In addition, our model could even automate the in-betweening process, such that users can easily create a temporally consistent animation by simply providing a character image as well as the start and end sketches. We will make the model public to facility the community.
****

DynPose: Largely Improving the Efficiency of Human Pose Estimation by a Simple Dynamic Framework
Yalong Xu, Lin Zhao, Chen Gong, Guangyu Li, Di Wang, Nannan Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1160-1169
Top-down approaches for human pose estimation (HPE) have reached a high level of sophistication, exemplified by models such as HRNet and ViTPose. Nonetheless, the low efficiency of top-down methods is a recognized issue that has not been sufficiently explored in current research. Our analysis suggests that the primary cause of inefficiency stems from the substantial diversity found in pose samples. On one hand, simple poses can be accurately estimated without requiring the computational resources of larger models. On the other hand, a more prominent issue arises from the abundance of bounding boxes, which remain excessive even after NMS. In this paper, we present a straightforward yet effective dynamic framework called DynPose, designed to match diverse pose samples with the most appropriate models, thereby ensuring optimal performance and high efficiency. Specifically, the framework contains a lightweight router and two pre-trained HPE models: one small and one large. The router is optimized to classify samples and dynamically determine the appropriate inference paths. Extensive experiments demonstrate the effectiveness of the framework. For example, using ResNet-50 and HRNet-W32 as the pre-trained models, our DynPose achieves an almost 50% increase in speed over HRNet-W32 while maintaining the same-level accuracy. More importantly, the framework can be generalized to other pre-trained models and datasets without re

-training or fine-tuning. Code is available at https://github.com/Aritoria/DynPose.
****

Arbitrary-steps Image Super-resolution via Diffusion Inversion
Zongsheng Yue, Kang Liao, Chen Change Loy; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23153-23163
This study presents a new image super-resolution (SR) technique based on diffusion inversion, aiming at harnessing the rich image priors encapsulated in large pre-trained diffusion models to improve SR performance. We design a Partial noise Prediction strategy to construct an intermediate state of the diffusion model, which serves as the starting sampling point. Central to our approach is a deep noise predictor to estimate the optimal noise maps for the forward diffusion process. Once trained, this noise predictor can be used to initialize the sampling process partially along the diffusion trajectory, generating the desirable high-resolution result. Compared to existing approaches, our method offers a flexible and efficient sampling mechanism that supports an arbitrary number of sampling steps, ranging from one to five. Even with a single sampling step, our method demonstrates superior or comparable performance to recent state-of-the-art approaches. The code and model will be made publicly available.
****

LiSu: A Dataset and Method for LiDAR Surface Normal Estimation
Dušan Mali■, Christian Fruhwirth-Reisinger, Samuel Schulter, Horst Possegger; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17039-17049
While surface normals are widely used to analyse 3D scene geometry, surface normal estimation from LiDAR point clouds remains severely underexplored. This is caused by the lack of large-scale annotated datasets on the one hand, and lack of methods that can robustly handle the sparse and often noisy LiDAR data in a reasonable time on the other hand. We address these limitations using a traffic simulation engine and present LiSu, the first large-scale, synthetic LiDAR point cloud dataset with ground truth surface normal annotations, eliminating the need for tedious manual labeling. Additionally, we propose a novel method that exploits the spatiotemporal characteristics of autonomous driving data to enhance surface normal estimation accuracy. By incorporating two regularization terms, we enforce spatial consistency among neighboring points and temporal smoothness across consecutive LiDAR frames. These regularizers are particularly effective in self-training settings, where they mitigate the impact of noisy pseudo-labels, enabling robust real-world deployment. We demonstrate the effectiveness of our method on LiSu, achieving state-of-the-art performance in LiDAR surface normal estimation. Moreover, we showcase its full potential in addressing the challenging task of synthetic-to-real domain adaptation, leading to improved neural surface reconstruction on real-world data.
****

Dynamic Neural Surfaces for Elastic 4D Shape Representation and Analysis
Awais Nizamani, Hamid Laga, Guanjin Wang, Farid Boussaid, Mohammed Bennamoun, Anuj Srivastava; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21783-21792
We propose a novel framework for the statistical analysis of genus-zero 4D surfaces, i.e., 3D surfaces that deform and evolve overtime. This problem is particularly challenging due to the arbitrary parameterizations of these surfaces and their varying deformation speeds, necessitating effective spatiotemporal registration. Traditionally, 4D surfaces are discretized, in space and time, before computing their spatiotemporal registrations, geodesics and statistics. However, this approach may result in suboptimal solutions and, as we demonstrate in this paper, is not necessary. In contrast, we treat 4D surfaces as continuous functions in both space and time. We introduce Dynamic Spherical Neural Surfaces (D-SNS), an efficient smooth and continuous spatiotemporal representation for genus-0 4D surfaces. We then demonstrate how to perform core 4D shape analysis tasks such as spatiotemporal registration, geodesics computation, and mean 4D shape estimation, directly on these continuous representations without upfront discretizati

on and meshing. By integrating neural representations with classical Riemannian geometry and statistical shape analysis techniques, we provide the building blocks for enabling full functional shape analysis. We demonstrate the efficiency of the framework on 4D human and face datasets.
****

Spk2SRImgNet: Super-Resolve Dynamic Scene from Spike Stream via Motion Aligned Collaborative Filtering
Yuanlin Wang, Yiyang Zhang, Ruiqin Xiong, Jing Zhao, Jian Zhang, Xiaopeng Fan, Tiejun Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11416-11426
Spike camera is a kind of neuromorphic camera that records dynamic scenes by firing a stream of binary spikes with extremely high temporal resolution. It demonstrates great potential for vision tasks in high-speed scenarios. One limitation in its current implementation is the relatively low spatial resolution. This paper develops a network called Spk2SRImgNet to super-resolve high resolution images from low resolution spike stream. However, fluctuations in spike stream hinder the performance of spike camera super resolution. To address this issue, we propose a motion aligned collaborative filtering (MACF) module, which is motivated by key ideas in classic image restoration schemes to mitigate fluctuations in spike data. MACF leverages the temporal similarity of spike stream to acquire similar features from neighboring moments via motion alignment. To separate disturbances from features, MACF filters these similar features jointly in transform domain to exploit representation sparsity, and generates refinement features that will be used to update initial fluctuated features. Specifically, MACF designs an inverse motion alignment operation to map these refinement features back to their original positions. The initial features are aggregated with the repositioned refinement features to enhance reliability. Experimental results demonstrate that the proposed method achieves state-of-the-art performance compared with existing methods.
****

ComfyBench: Benchmarking LLM-based Agents in ComfyUI for Autonomously Designing Collaborative AI Systems
Xiangyuan Xue, Zeyu Lu, Di Huang, Zidong Wang, Wanli Ouyang, Lei Bai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24614-24624
Much previous AI research has focused on developing monolithic models to maximize their intelligence, with the primary goal of enhancing performance on specific tasks. In contrast, this work attempts to study using LLM-based agents to design collaborative AI systems autonomously. To explore this problem, we first introduce ComfyBench to evaluate agents's ability to design collaborative AI systems in ComfyUI. ComfyBench is a comprehensive benchmark comprising 200 diverse tasks covering various instruction-following generation challenges, along with detailed annotations for 3,205 nodes and 20 workflows. Based on ComfyBench, we further develop ComfyAgent, a novel framework that empowers LLM-based agents to autonomously design collaborative AI systems by generating workflows. ComfyAgent is based on two core concepts. First, it represents workflows with code, which can be reversibly converted into workflows and executed as collaborative systems by the interpreter. Second, it constructs a multi-agent system that cooperates to learn from existing workflows and generate new workflows for a given task. While experimental results demonstrate that ComfyAgent achieves a comparable resolve rate to o1-preview and significantly surpasses other agents on ComfyBench, ComfyAgent has resolved only 15% of creative tasks. LLM-based agents still have a long way to go in autonomously designing collaborative AI systems. Progress with ComfyBench is paving the way for more intelligent and autonomous collaborative AI systems. Our code is available at: https://github.com/xxyQwQ/ComfyBench.
****

VideoGLaMM : A Large Multimodal Model for Pixel-Level Visual Grounding in Videos
Shehan Munasinghe, Hanan Gani, Wenqi Zhu, Jiale Cao, Eric Xing, Fahad Shahbaz Khan, Salman Khan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19036-19046

Fine-grained alignment between videos and text is challenging due to complex spatial and temporal dynamics in videos. Existing video-based Large Multimodal Models (LMMs) handle basic conversations but struggle with precise pixel-level grounding in videos. To address this, we introduce VideoGLaMM, a LMM designed for fine-grained pixel-level grounding in videos based on user-provided textual inputs. Our design seamlessly connects three key components: a Large Language Model, a dual vision encoder that emphasizes both spatial and temporal details, and a spatio-temporal decoder for accurate mask generation. This connection is facilitated via tunable V-L and L-V adapters that enable close Vision-Language (VL) alignment. The architecture is trained to synchronize both spatial and temporal elements of video content with textual instructions. To enable fine-grained grounding, we curate a multimodal dataset featuring detailed visually-grounded conversations using a semiautomatic annotation pipeline, resulting in a diverse set of 38k video-QA triplets along with 83k objects and 671k masks. We evaluate VideoGLaMM on three challenging tasks: Grounded Conversation Generation, Visual Grounding, and Referring Video Segmentation. Experimental results show that our model consistently outperforms existing approaches across all three tasks.
****

Incomplete Multi-View Multi-label Learning via Disentangled Representation and Label Semantic Embedding

Xu Yan, Jun Yin, Jie Wen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30722-30731

In incomplete multi-view multi-label learning scenarios, it is crucial to use the incomplete multi-view data to extract consistent and specific representations from different data sources and to fully exploit the missing label information. However, most previous approaches ignore the separation problem between view-shared and specific information. To address this problem, in this paper, we propose a method that can separate view-consistent features from view-specific features under the Variational Autoencoder (VAE) framework. Specifically, we first introduce cross-view reconstruction to capture view-consistent features and extract shared information from different views through unsupervised pre-training. Subsequently, we develop a disentangling module to learn specific features by minimizing the variational upper bound of mutual information between consistent and specific features. Finally, we utilize prior label relevance information derived from training data to guide the learning of the distribution of label semantic embeddings, aggregating relevant semantic embeddings and maintaining the label relevance topology in the semantic space. In extensive experiments, our model outperforms existing state-of-the-art algorithms on several real-world datasets, which fully validates its strong adaptability to missing views and labels.
****

AutoURDF: Unsupervised Robot Modeling from Point Cloud Frames Using Cluster Registration

Jiong Lin, Lechen Zhang, Kwansoo Lee, Jialong Ning, Judah Goldfeder, Hod Lipson; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27628-27637

Robot description models are essential for simulation and control, yet their creation often requires significant manual effort. To streamline this modeling process, we introduce AutoURDF, an unsupervised approach for constructing description files for unseen robots from point cloud frames. Our method leverages a cluster-based point cloud registration model that tracks the 6-DoF transformations of point clusters. Through analyzing cluster movements, we hierarchically address the following challenges: (1) moving part segmentation, (2) body topology inference, and (3) joint parameter estimation. The complete pipeline produces robot description files that are fully compatible with existing simulators. We validate our method across a variety of robots, using both synthetic and real-world scan data. Results indicate that our approach outperforms previous methods in registration and body topology estimation accuracy, offering a scalable solution for automated robot modeling.
****

ZeroGrasp: Zero-Shot Shape Reconstruction Enabled Robotic Grasping

Shun Iwase, Muhammad Zubair Irshad, Katherine Liu, Vitor Guizilini, Robert Lee, Takuya Ikeda, Ayako Amma, Koichi Nishiwaki, Kris Kitani, Rares Ambrus, Sergey Zakharov; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17405-17415

Robotic grasping is a cornerstone capability of embodied systems. Many methods directly output grasps from partial information without modeling the geometry of the scene, leading to suboptimal motion and even collisions. To address these issues, we introduce ZeroGrasp, a novel framework that simultaneously performs 3D reconstruction and grasp pose prediction in near real-time. A key insight of our method is that occlusion reasoning and modeling the spatial relationships between objects is beneficial for both accurate reconstruction and grasping. We couple our method with a novel large-scale synthetic dataset, which comprises 1M photo-realistic images, high-resolution 3D reconstructions and 11.3B physically-valid grasp pose annotations for 12K objects from the Objaverse-LVIS dataset. We evaluate ZeroGrasp on the GraspNet-1B benchmark as well as through real-world robot experiments. ZeroGrasp achieves state-of-the-art performance and generalizes to novel real-world objects by leveraging synthetic data.

****

Golden Cudgel Network for Real-Time Semantic Segmentation

Guoyu Yang, Yuan Wang, Daming Shi, Yanzhong Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25367-25376

Recent real-time semantic segmentation models, whether single-branch or multi-branch, achieve good performance and speed. However, their speed is limited by multi-path blocks, and some depend on high-performance teacher models for training. To overcome these issues, we propose Golden Cudgel Network (GCNet). Specifically, GCNet uses vertical multi-convolutions and horizontal multi-paths for training, which are reparameterized into a single convolution for inference, optimizing both performance and speed. This design allows GCNet to self-enlarge during training and self-contract during inference, effectively becoming a "teacher model" without needing external ones. Experimental results show that GCNet outperforms existing state-of-the-art models in terms of performance and speed on the Cityscapes, CamVid, and Pascal VOC 2012 datasets. The code is available at https://github.com/gyyang23/GCNet.

****

PDFactor: Learning Tri-Perspective View Policy Diffusion Field for Multi-Task Robotic Manipulation

Jingyi Tian, Le Wang, Sanping Zhou, Sen Wang, Jiayi Li, Haowen Sun, Wei Tang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15757-15767

Robotic manipulation based on visual observations and natural language instructions is a long-standing challenge in robotics. Yet prevailing approaches model action distribution by adopting explicit or implicit representations, which often struggle to achieve a trade-off between accuracy and efficiency. In response, we propose PDFactor, a novel framework that models action distribution with a hybrid triplane representation. In particular, PDFactor decomposes 3D point cloud into three orthogonal feature planes and leverages a tri-perspective view transformer to produce dense cubic features as a latent diffusion field aligned with observation space representing 6-DoF action probability distribution at an arbitrary location. We employ a small denoising network conceptually as both a parameterized loss function measuring the quality of the learned latent features and an action gradient decoder to sample actions from the latent diffusion field during inference. This design enables our PDFactor to benefit from spatial awareness of explicit representation and arbitrary resolution of implicit representation, rendering it with manipulation accuracy, inference efficiency, and model scalability. Experiments demonstrate that PDFactor outperforms state-of-the-art approaches across a diverse range of manipulation tasks in RLBench simulation. Moreover, PDFactor can effectively learn multi-task policies from a limited number of human demonstrations, achieving promising accuracy in a variety of real-world manipulation tasks.

****

VideoTree: Adaptive Tree-based Video Representation for LLM Reasoning on Long Videos
Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, Mohit Bansal; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3272-3283
Long-form video understanding has been a challenging task due to the high redundancy in video data and the abundance of query-irrelevant information. To tackle this challenge, we propose VideoTree, a training-free framework which builds a query-adaptive and hierarchical video representation for LLM reasoning over long-form videos. First, VideoTree extracts query-relevant information from the input video through an iterative process, progressively refining the selection of key frames based on their relevance to the query. Furthermore, VideoTree leverages the inherent hierarchical structure of long video data, which is often overlooked by existing LLM-based methods. Specifically, we incorporate multi-granularity information into a tree-based representation, allowing VideoTree to extract query-relevant details from long videos in a coarse-to-fine manner. This enables the model to effectively handle a wide range of video queries with varying levels of detail. Finally, VideoTree aggregates the hierarchical query-relevant information within the tree structure and feeds it into an LLM reasoning model to answer the query. Our experiments show that VideoTree improves both reasoning accuracy and efficiency compared to existing methods. Specifically, VideoTree outperforms the existing training-free approaches on the popular EgoSchema and NExT-QA benchmarks with less inference time, achieving 61.1% and 75.6% accuracy on the test set without additional video-specific training. Moreover, on the long split of Video-MME benchmark (average 44 minutes), the training-free VideoTree framework achieves better performance than the strong proprietary GPT-4V model and many other MLLMs that were extensively trained on video data. Our code is provided in the supplementary materials and will be made public.
****
Multi-modal Contrastive Learning with Negative Sampling Calibration for Phenotypic Drug Discovery
Jiahua Rao, Hanjing Lin, Leyu Chen, Jiancong Xie, Shuangjia Zheng, Yuedong Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30752-30762
Phenotypic drug discovery presents a promising strategy for identifying first-in-class drugs by bypassing the need for specific drug targets. Recent advances in cell-based phenotypic screening tools, including Cell Painting and the LINCS L1000, provide essential cellular data that capture biological responses to compounds. While the integration of the multi-modal data enhances the use of contrastive learning (CL) methods for molecular phenotypic representation, these approaches treat all negative pairs equally, failing to discriminate molecules with similar phenotypes. To address these challenges, we introduce a foundational framework MINER that dynamically estimates the likelihoods of sample pairs as negative pairs based on uni-modal disentangled representations. In addition, our approach incorporates a mixture fusion strategy to effectively integrate multimodal data, even in cases where certain modalities are missing. Extensive experiments demonstrate that our method enhances both molecular property prediction and molecule-phenotype retrieval accuracy. Moreover, it successfully recommends drug candidates from phenotype for complex diseases documented in the literature. These findings underscore MINER's potential to advance drug discovery by enabling deeper insights into disease mechanisms and improving drug candidate recommendations.
****
R-TPT: Improving Adversarial Robustness of Vision-Language Models through Test-Time Prompt Tuning
Lijun Sheng, Jian Liang, Zilei Wang, Ran He; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29958-29967
Vision-language models (VLMs), such as CLIP, have gained significant popularity as foundation models, with numerous fine-tuning methods developed to enhance performance on downstream tasks. However, due to their inherent vulnerability and the common practice of selecting from a limited set of open-source models, VLMs s

uffer from a higher risk of adversarial attacks than traditional vision models. Existing defense techniques typically rely on adversarial fine-tuning during training, which requires labeled data and lacks of flexibility for downstream tasks. To address these limitations, we propose robust test-time prompt tuning (R-TPT), which mitigates the impact of adversarial attacks during the inference stage. We first reformulate the classic marginal entropy objective by eliminating the term that introduces conflicts under adversarial conditions, retaining only the pointwise entropy minimization. Furthermore, we introduce a plug-and-play reliability-based weighted ensembling strategy, which aggregates useful information from reliable augmented views to strengthen the defense. R-TPT enhances defense against adversarial attacks without requiring labeled training data while offering high flexibility for inference tasks. Extensive experiments on widely used benchmarks with various attacks demonstrate the effectiveness of R-TPT.

****

Distinguish Then Exploit: Source-free Open Set Domain Adaptation via Weight Barcode Estimation and Sparse Label Assignment

Weiming Liu, Jun Dan, Fan Wang, Xinting Liao, Junhao Dong, Hua Yu, Shunjie Dong, Lianyong Qi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4927-4938

Nowadays, domain adaptation techniques have been widely investigated for knowledge sharing from labeled source domain to unlabeled target domain. However, target domain may include some data samples that belong to unknown categories in real-world scenarios. Moreover, the target domain cannot access the source data samples due to privacy-preserving restrictions. In this paper, we focus on the source-free open set domain adaptation problem which includes two main challenges, i.e.,how to distinguish known and unknown target samples and how to exploit useful source information to provide trustworthy pseudo labels for known target samples. Existing approaches that directly applying conventional domain alignment methods could lead to sample mismatch and misclassification in this scenario.To overcome these issues, we propose Distinguish Then Exploit model (DTE) with two components, i.e., weight barcode estimation and sparse label assignment. Weight barcode estimation first calculate marginal probability of target samples via partially unbalanced optimal transport then quantize barcode results to distinguish unknown target sample. Sparse label assignment utilizes sparse sample-label matching via proximal term to fully exploit useful source information. Our empirically study on several datasets shows that DTE outperforms the state-of-the-art models on tackling the source-free open set domain adaptation problem.

****

Multi-Sensor Object Anomaly Detection: Unifying Appearance, Geometry, and Internal Properties

Wenqiao Li, Bozhong Zheng, Xiaohao Xu, Jinye Gan, Fading Lu, Xiang Li, Na Ni, Zheng Tian, Xiaonan Huang, Shenghua Gao, Yingna Wu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9984-9993

Object anomaly detection is essential for industrial quality inspection, yet traditional single-sensor methods face critical limitations. They fail to capture the wide range of anomaly types, as single sensors are often constrained to either external appearance, geometric structure, or internal properties. To overcome these challenges, we introduce MulSen-AD, the first high-resolution, multi-sensor anomaly detection dataset tailored for industrial applications. MulSen-AD unifies data from RGB cameras, laser scanners, and lock-in infrared thermography, effectively capturing external appearance, geometric deformations, and internal defects. The dataset spans 15 industrial products with diverse, real-world anomalies. We also present MulSen-AD Bench, a benchmark designed to evaluate multi-sensor methods, and propose MulSen-TripleAD, a decision-level fusion algorithm that integrates these three modalities for robust, unsupervised object anomaly detection. Our experiments demonstrate that multi-sensor fusion substantially outperforms single-sensor approaches, achieving 96.1% AUROC in object-level detection accuracy. These results highlight the importance of integrating multi-sensor data for comprehensive industrial anomaly detection. The dataset and code are available at https://github.com/ZZZBBBZZZ/MulSen-AD to support further research.

****
SplatFlow: Multi-View Rectified Flow Model for 3D Gaussian Splatting Synthesis
Hyojun Go, Byeongjun Park, Jiho Jang, Jin-Young Kim, Soonwoo Kwon, Changick Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21524-21536
Text-based generation and editing of 3D scenes hold significant potential for streamlining content creation through intuitive user interactions. While recent advances leverage 3D Gaussian Splatting (3DGS) for high-fidelity and real-time rendering, existing methods are often specialized and task-focused, lacking a unified framework for both generation and editing. In this paper, we introduce SplatFlow, a comprehensive framework that addresses this gap by enabling direct 3DGS generation and editing. SplatFlow comprises two main components: a multi-view rectified flow (RF) model and a Gaussian Splatting Decoder (GSDecoder). The multi-view RF model operates in latent space, generating multi-view images, depths, and camera poses simultaneously, conditioned on text prompts--thus addressing challenges like diverse scene scales and complex camera trajectories in real-world settings. Then, the GSDecoder efficiently translates these latent outputs into 3DGS representations through a feed-forward 3DGS method. Leveraging training-free inversion and inpainting techniques, SplatFlow enables seamless 3DGS editing and supports a broad range of 3D tasks--including object editing, novel view synthesis, and camera pose estimation--within a unified framework without requiring additional complex pipelines. We validate SplatFlow's capabilities on the MVImgNet and DL3DV-7K datasets, demonstrating its versatility and effectiveness in various 3D generation, editing, and inpainting-based tasks.
****
Fancy123: One Image to High-Quality 3D Mesh Generation via Plug-and-Play Deformation
Qiao Yu, Xianzhi Li, Yuan Tang, Xu Han, Long Hu, Yixue Hao, Min Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 595-604
Generating 3D meshes from a single image is an important but ill-posed task. Existing methods mainly adopt 2D multiview diffusion models to generate intermediate multiview images, and use the Large Reconstruction Model (LRM) to create the final meshes. However, the multiview images exhibit local inconsistencies, and the meshes often lack fidelity to the input image or look blurry. We propose Fancy123, featuring two enhancement modules and an unprojection operation to address the above three issues, respectively. The appearance enhancement module deforms the 2D multiview images to realign misaligned pixels for better multiview consistency. The fidelity enhancement module deforms the 3D mesh to match the input image. The unprojection of the input image and deformed multiview images onto LRM's generated mesh ensures high clarity, discarding LRM's predicted blurry-looking mesh colors. Extensive qualitative and quantitative experiments verify Fancy123's SoTA performance with significant improvement. Also, the two enhancement modules are plug-and-play and work at inference time, allowing seamless integration into various existing single-image-to-3D methods. Project page: https://github.com/YuQiao0303/Fancy123.
****
Dense Dispersed Structured Light for Hyperspectral 3D Imaging of Dynamic Scenes
Suhyun Shin, Seungwoo Yoon, Ryota Maeda, Seung-Hwan Baek; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16589-16598
Hyperspectral 3D imaging captures both depth maps and hyperspectral images, enabling comprehensive geometric and material analysis. Recent methods achieve high spectral and depth accuracy; however, they require long acquisition times--often over several minutes--or rely on large, expensive systems, restricting their use to static scenes. We present Dense Dispersed Structured Light (DDSL), an accurate hyperspectral 3D imaging method for dynamic scenes that utilizes stereo RGB cameras and an RGB projector equipped with an affordable diffraction grating film. We design spectrally multiplexed DDSL patterns that significantly reduce the number of required projector patterns, thereby accelerating acquisition speed. Additionally, we formulate an image formation model and a reconstruction method t

o estimate a hyperspectral image and depth map from captured stereo images. As t
he first practical and accurate hyperspectral 3D imaging method for dynamic scen
es, we experimentally demonstrate that DDSL achieves a spectral resolution of 15
.5 nm full width at half maximum (FWHM), 4 mm depth error, and 6.6 fps.
****

PlanarSplatting: Accurate Planar Surface Reconstruction in 3 Minutes
Bin Tan, Rui Yu, Yujun Shen, Nan Xue; Proceedings of the Computer Vision and Pat
tern Recognition Conference (CVPR), 2025, pp. 1190-1199
This paper presents PlanarSplatting, an ultra-fast and accurate surface reconstr
uction approach for multi-view indoor images. We take the 3D planes as the main
objective due to their compactness and structural expressiveness in indoor scene
s, and develop an explicit optimization framework that learns to fit the expecte
d surface of indoor scenes by splatting the 3D planes into 2.5D depth and normal
 maps. As our PlanarSplatting operates directly on the 3D plane primitives, it e
liminates the dependencies on 2D/3D plane detection and plane matching/tracking
for planar surface reconstruction. Furthermore, with the essential merits of pla
ne-based representation coupled with CUDA-based implementation of planar splatti
ng functions, PlanarSplatting reconstructs an indoor scene in 3 minutes while ha
ving significantly better geometric accuracy. Thanks to our ultra-fast reconstru
ction speed, the largest quantitative evaluation on the ScanNet and ScanNet++ da
tasets over hundreds of scenes clearly demonstrated the advantages of our method
. We believe that our accurate and ultrafast planar surface reconstruction metho
d will be applied in the structured data curation for surface reconstruction in
the future. The code of our CUDA implementation will be publicly available.
****

Omni-ID: Holistic Identity Representation Designed for Generative Tasks
Guocheng Qian, Kuan-Chieh Wang, Or Patashnik, Negin Heravi, Daniil Ostashev, Ser
gey Tulyakov, Daniel Cohen-Or, Kfir Aberman; Proceedings of the Computer Vision
and Pattern Recognition Conference (CVPR), 2025, pp. 8786-8795
We introduce Omni-ID, a novel facial representation designed specifically for ge
nerative tasks. Omni-ID encodes holistic information about an individual's appea
rance across diverse expressions and poses within a fixed-size representation. I
t consolidates information from a varied number of unstructured input images int
o a structured representation, where each entry represents certain global or loc
al identity features. Our approach uses a few-to-many identity reconstruction tr
aining paradigm, where a limited set of input images is used to reconstruct mult
iple target images of the same individual in various poses and expressions. A mu
lti-decoder framework is introduced to leverage the complementary strengths of d
iverse decoders during training. Unlike conventional representations, such as Ar
cFace and CLIP, which are typically learned through discriminative or contrastiv
e objectives, Omni-ID is optimized with a generative objective, resulting in a m
ore comprehensive and nuanced identity capture for generative tasks. Trained on
our MFHQ dataset -- a multi-view facial image collection, Omni-ID demonstrates s
ubstantial improvements over conventional representations across various generat
ive tasks.
****

MM-OR: A Large Multimodal Operating Room Dataset for Semantic Understanding of H
igh-Intensity Surgical Environments
Ege Özsoy, Chantal Pellegrini, Tobias Czempiel, Felix Tristram, Kun Yuan, David
Bani-Harouni, Ulrich Eck, Benjamin Busam, Matthias Keicher, Nassir Navab; Procee
dings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp
. 19378-19389
Operating rooms (ORs) are complex, high-stakes environments requiring precise un
derstanding of interactions among medical staff, tools, and equipment for enhanc
ing surgical assistance, situational awareness, and patient safety. Current data
sets fall short in scale, realism and do not capture the multimodal nature of OR
 scenes, limiting progress in OR modeling. To this end, we introduce MM-OR, a re
alistic and large-scale multimodal spatiotemporal OR dataset, and the first data
set to enable multimodal scene graph generation. MM-OR captures comprehensive OR
 scenes containing RGB-D data, detail views, audio, speech transcripts, robotic

logs, and tracking data and is annotated with panoptic segmentations, semantic s
cene graphs, and downstream task labels. Further, we propose MM2SG, the first mu
ltimodal large vision-language model for scene graph generation, and through ext
ensive experiments, demonstrate its ability to effectively leverage multimodal i
nputs. Together, MM-OR and MM2SG establish a new benchmark for holistic OR under
standing, and open the path towards multimodal scene analysis in complex, high-s
takes environments. We will publish all our code and dataset upon acceptance.
****

MIRE: Matched Implicit Neural Representations
Dhananjaya Jayasundara, Heng Zhao, Demetrio Labate, Vishal M. Patel; Proceedings
 of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 827
9-8288
Implicit Neural Representations (INRs) are continuous function learners for conv
entional digital signal representations. With the aid of positional embeddings a
nd/or exhaustively fine-tuned activation functions, INRs have surpassed many lim
itations of traditional discrete representations. However, existing works only f
ind a continuous representation for the digital signal by solely using a single,
 fixed activation function throughout the INR, and it has not yet been explored
to match the INR to the given signal. As current INRs are not matched to the sig
nal being represented by the INR, we hypothesize that this approach could restri
ct the representation power and generalization capabilities of INRs, limiting th
eir broader applicability. A way to match the INR to the signal being represente
d is through matching the activation of each layer in the sense of minimizing th
e mean squared loss. In this paper, we introduce MIRE, a method to find the high
ly matched activation function for each layer in INR through dictionary learning
. To showcase the effectiveness of the proposed method, we utilize a dictionary
that includes seven activation atoms: Raised Cosines (RC), Root Raised Cosines (
RRC), Prolate Spheroidal Wave Function (PSWF), Sinc, Gabor Wavelet, Gaussian, an
d Sinusoidal. Experimental results demonstrate that MIRE not only significantly
improves INR performance across various tasks, such as image representation, ima
ge inpainting, 3D shape representation, novel view synthesis, super-resolution,
and reliable edge detection, but also eliminates the need for the previously req
uired exhaustive search for activation parameters, which had to be conducted eve
n before INR training could begin.
****

AeSPa : Attention-guided Self-supervised Parallel Imaging for MRI Reconstruction
Jinho Joo, Hyeseong Kim, Hyeyeon Won, Deukhee Lee, Taejoon Eo, Dosik Hwang; Proc
eedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025,
pp. 5217-5226
This study introduces a novel zero-shot scan-specific self-supervised reconstruc
tion method for magnetic resonance imaging (MRI) to reduce scan times. Conventio
nal supervised reconstruction methods require large amounts of fully-sampled ref
erence data, which is often impractical to obtain and can lead to artifacts by o
verly emphasizing learned patterns. Existing zero-shot scan-specific methods hav
e attempted to overcome this data dependency but show limited performance due to
 insufficient utilization of k-space information and constraints derived from MR
I forward model. To address these limitations, we introduce a framework utilizin
g all acquired k-space measurements for both network inputs and training targets
. While this framework suffers from training instability, we resolve these chall
enges through three key components: an Attention-guided K-space Selective Mechan
ism (AKSM) that provides indirect constraints for non-sampled k-space points, It
eration-wise K-space Masking (IKM) that enhances training stability, and a robus
t sensitivity map estimation model utilizing cross-channel constraint that perfo
rms effectively even at high reduction factors. Experimental results on the Fast
MRI knee and brain datasets with reduction factors of 4 and 8 demonstrate that t
he proposed method achieves superior reconstruction quality and faster convergen
ce compared to existing zero-shot scan-specific methods, making it suitable for
practical clinical applications. The implementation of our proposed method is pu
blicly available at https://github.com/joojinho97/AeSPa.git.
****

Boltzmann Attention Sampling for Image Analysis with Small Objects
Theodore Zhao, Sid Kiblawi, Naoto Usuyama, Ho Hin Lee, Sam Preston, Hoifung Poon , Mu Wei; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25950-25959
Detecting and segmenting small objects, such as lung nodules and tumor lesions, remains a critical challenge in image analysis. These objects often occupy less than 0.1% of an image, making traditional transformer architectures inefficient and prone to performance degradation due to redundant attention computations on irrelevant regions. Existing sparse attention mechanisms rely on rigid hierarchi cal structures, which are poorly suited for detecting small, variable, and uncer tain object locations. In this paper, we propose BoltzFormer, a novel transforme r-based architecture designed to address these challenges through dynamic sparse attention. BoltzFormer identifies and focuses attention on relevant areas by mo deling uncertainty using a Boltzmann distribution with an annealing schedule. In itially, a higher temperature allows broader area sampling in early layers, when object location uncertainty is greatest. As the temperature decreases in later layers, attention becomes more focused, enhancing efficiency and accuracy. Boltz Former seamlessly integrates into existing transformer architectures via a modul ar Boltzmann attention sampling mechanism. Comprehensive evaluations on benchmar k datasets demonstrate that BoltzFormer significantly improves segmentation perf ormance for small objects while reducing attention computation by an order of ma gnitude compared to previous state-of-the-art methods.
****
Dora: Sampling and Benchmarking for 3D Shape Variational Auto-Encoders
Rui Chen, Jianfeng Zhang, Yixun Liang, Guan Luo, Weiyu Li, Jiarui Liu, Xiu Li, X iaoxiao Long, Jiashi Feng, Ping Tan; Proceedings of the Computer Vision and Patt ern Recognition Conference (CVPR), 2025, pp. 16251-16261
Recent 3D content generation pipelines commonly employ Variational Autoencoders (VAEs) to encode shapes into compact latent representations for diffusion-based generation. However, the widely adopted uniform point sampling strategy in Shape VAE training often leads to a significant loss of geometric details, limiting the quality of shape reconstruction and downstream generation tasks. We present Dora-VAE, a novel approach that enhances VAE reconstruction through our proposed sharp edge sampling strategy and a dual cross-attention mechanism. By identifyi ng and prioritizing regions with high geometric complexity during training, our method significantly improves the preservation of fine-grained shape features. S uch sampling strategy and the dual attention mechanism enable the VAE to focus o n crucial geometric details that are typically missed by uniform sampling approa ches. To systematically evaluate VAE reconstruction quality, we additionally pro pose Dora-bench, a benchmark that quantifies shape complexity through the densit y of sharp edges, introducing a new metric focused on reconstruction accuracy at these salient geometric features. Extensive experiments on the Dora-bench demon strate that Dora-VAE achieves comparable reconstruction quality to the state-of- the-art dense XCube-VAE while requiring a latent space at least 8xsmaller (1,280 vs. > 10,000 codes). Project page: https://aruichen.github.io/Dora/.
****
Generalized Recorrupted-to-Recorrupted: Self-Supervised Learning Beyond Gaussian Noise
Brayan Monroy, Jorge Bacca, Julián Tachella; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28155-28164
Recorrupted-to-Recorrupted (R2R) has emerged as a methodology for training deep networks for image restoration in a self-supervised manner from noisy measuremen t data alone, demonstrating equivalence in expectation to the supervised squared loss in the case of Gaussian noise. However, its effectiveness with non-Gaussia n noise remains unexplored. In this paper, we propose Generalized R2R (GR2R), ex tending the R2R framework to handle a broader class of noise distribution as add itive noise like log-Rayleigh and address the natural exponential family includi ng Poisson, Gamma and Binomial noise distributions, which play a key role in man y applications including low-photon imaging and synthetic aperture radar. We sho w that the GR2R loss is an unbiased estimator of the supervised loss and that th

e popular Stein's unbiased risk estimator can be seen as a special case. A serie s of experiments with Gaussian, Poisson, and Gamma noise validate GR2R's perform ance, showing its effectiveness compared to other self-supervised methods.
****

Once-Tuning-Multiple-Variants: Tuning Once and Expanded as Multiple Vision-Langu age Model Variants
Chong Yu, Tao Chen, Zhongxue Gan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14712-14722
Vision-language model (VLM) is one of the most important models for multi-modal tasks. Real industrial applications often meet the challenge of adapting VLMs to different scenarios, such as varying hardware platforms or performance requirem ents. Traditional methods involve training or fine-tuning to adapt multiple uniq ue VLMs or using model compression techniques to create multiple compact models. These approaches are complex and resource-intensive. This paper introduces a no vel paradigm called Once-Tuning-Multiple-Variants (OTMV). OTMV requires only a s ingle tuning process to inject dynamic weight expansion capacity into the VLM wi th dynamic expansion capacity. This tuned VLM can then be expanded into multiple variants tailored for different scenarios in inference. The tuning mechanism of OTMV is inspired by the mathematical series expansion theorem, which helps to r educe the parameter size and memory requirements while maintaining accuracy for VLM. Experiment results show that OTMV-tuned models achieve comparable accuracy to baseline VLMs across various visual-language tasks. The experiments also demo nstrate the dynamic expansion capability of OTMV-tuned VLMs, outperforming tradi tional model compression and adaptation techniques in terms of accuracy and effi ciency.
****

Dynamic Motion Blending for Versatile Motion Editing
Nan Jiang, Hongjie Li, Ziye Yuan, Zimo He, Yixin Chen, Tengyu Liu, Yixin Zhu, Si yuan Huang; Proceedings of the Computer Vision and Pattern Recognition Conferenc e (CVPR), 2025, pp. 22735-22745
Text-guided motion editing enables high-level semantic control and iterative mod ifications beyond traditional keyframe animation. Existing methods rely on limit ed pre-collected training triplets (original motion, edited motion, and instruct ion), which severely hinders their versatility in diverse editing scenarios. We introduce MotionCutMix, an online data augmentation technique that dynamically g enerates training triplets by blending body part motions based on input text. Wh ile MotionCutMix effectively expands the training distribution, the compositiona l nature introduces increased randomness and potential body part incoordination. To model such a rich distribution, we present MotionReFit, an auto-regressive d iffusion model with a motion coordinator. The auto-regressive architecture facil itates learning by decomposing long sequences, while the motion coordinator miti gates the artifacts of motion composition. Our method handles both spatial and t emporal motion edits directly from high-level human instructions, without relyin g on additional specifications or LLM. Through extensive experiments, we show th at MotionReFit achieves state-of-the-art performance in text-guided motion editi ng. Ablation studies further verify that MotionCutMix significantly improves the model's generalizability while maintaining training convergence.
****

StdGEN: Semantic-Decomposed 3D Character Generation from Single Images
Yuze He, Yanning Zhou, Wang Zhao, Zhongkai Wu, Kaiwen Xiao, Wei Yang, Yong-Jin L iu, Xiao Han; Proceedings of the Computer Vision and Pattern Recognition Confere nce (CVPR), 2025, pp. 26345-26355
We present StdGEN, an innovative pipeline for generating semantically decomposed high-quality 3D characters from single images, enabling broad applications in v irtual reality, gaming, and filmmaking, etc. Unlike previous methods which strug gle with limited decomposability, unsatisfactory quality, and long optimization times, StdGEN features decomposability, effectiveness and efficiency; i.e., it g enerates intricately detailed 3D characters with separated semantic components s uch as the body, clothes, and hair, in three minutes. At the core of StdGEN is o ur proposed Semantic-aware Large Reconstruction Model (S-LRM), a transformer-bas

ed generalizable model that jointly reconstructs geometry, color and semantics f
rom multi-view images in a feed-forward manner. A differentiable multi-layer sem
antic surface extraction scheme is introduced to acquire meshes from hybrid impl
icit fields reconstructed by our S-LRM. Additionally, a specialized efficient mu
lti-view diffusion model and an iterative multi-layer surface refinement module
are integrated into the pipeline to facilitate high-quality, decomposable 3D cha
racter generation. Extensive experiments demonstrate our state-of-the-art perfor
mance in 3D anime character generation, surpassing existing baselines by a signi
ficant margin in geometry, texture and decomposability. StdGEN offers ready-to-u
se semantic-decomposed 3D characters and enables flexible customization for a wi
de range of applications. Project page: https://stdgen.github.io
****

Reconstructing Animals and the Wild
Peter Kulits, Michael J. Black, Silvia Zuffi; Proceedings of the Computer Vision
and Pattern Recognition Conference (CVPR), 2025, pp. 16565-16577
The notion of 3D reconstruction as scene understanding is foundational in comput
er vision. Reconstructing 3D scenes from 2D visual observations necessitates str
ong priors to disambiguate structure. Much work has been focused on the anthropo
centric, which, characterized by smooth surfaces, coherent normals, and regular
edges, allows for the integration of strong geometric inductive biases. Here, we
consider a more challenging problem where such assumptions do not hold: the rec
onstruction of natural scenes containing trees, bushes, boulders, and animals. W
hile numerous works have attempted to tackle the problem of reconstructing anima
ls in the wild, they have focused solely on the animal, neglecting environmental
context. This limits their usefulness for analysis tasks, as animals exist inhe
rently within the 3D world, and information is lost when environmental factors a
re disregarded. We propose a method to reconstruct natural scenes from single im
ages. We base our approach on recent advances leveraging the strong world priors
ingrained in Large Language Models and train an autoregressive model to decode
a CLIP embedding into a structured, compositional scene representation, encompas
sing both animals and the wild. To enable this, we propose a synthetic dataset c
omprising one million images and thousands of assets. Through our introduction o
f a CLIP-projection head, we demonstrate that our approach generalizes to the ta
sk of reconstructing animals and their environments in real-world images, despit
e having been trained solely on synthetic data. We release our dataset and code
to encourage future research at https://raw.is.tue.mpg.de/
****

RobSense: A Robust Multi-modal Foundation Model for Remote Sensing with Static,
Temporal, and Incomplete Data Adaptability
Minh Kha Do, Kang Han, Phu Lai, Khoa T. Phan, Wei Xiang; Proceedings of the Comp
uter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7427-7436
Foundation models for remote sensing have garnered increasing attention for thei
r strong performance across various observation tasks. However, current models l
ack robustness in managing diverse input types and handling incomplete data in d
ownstream tasks. In this paper, we propose RobSense, a robust multi-modal founda
tion model for Multi-spectral and Synthetic Aperture Radar data. RobSense is des
igned with modular components and pre-trained by a combination of temporal multi
-modal alignment and masked autoencoder strategies on a huge-scale dataset. Ther
efore, it can effectively support diverse input types, from static to temporal,
uni-modal to multi-modal. To further handle the incomplete data, we incorporate
two uni-modal latent reconstructors that recover rich representations from incom
plete inputs, addressing variability in spectral bands and temporal sequence irr
egularities. Extensive experiments demonstrate that RobSense consistently outper
forms state-of-the-art baselines on complete datasets across four input types fo
r segmentation, classification, and change detection. On incomplete datasets, Ro
bSense outperforms the baselines by considerably larger margins when the missing
rate increases. Project page: https://ikhado.github.io/robsense/
****

Spatiotemporal Decoupling for Efficient Vision-Based Occupancy Forecasting
Jingyi Xu, Xieyuanli Chen, Junyi Ma, Jiawei Huang, Jintao Xu, Yue Wang, Ling Pei

; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22338-22347
The task of occupancy forecasting (OCF) involves utilizing past and present perception data to predict future occupancy states of autonomous vehicle surrounding environments, which is critical for downstream tasks such as obstacle avoidance and path planning. Existing 3D OCF approaches struggle to predict plausible spatial details for movable objects and suffer from slow inference speeds due to neglecting the bias and uneven distribution of changing occupancy states in both space and time. In this paper, we propose a novel spatiotemporal decoupling vision-based paradigm to explicitly tackle the bias and achieve both effective and efficient 3D OCF. To tackle spatial bias in empty areas, we introduce a novel spatial representation that decouples the conventional dense 3D format into 2D bird's-eye view (BEV) occupancy with corresponding height values, enabling 3D OCF derived only from 2D predictions thus enhancing efficiency. To reduce temporal bias on static voxels, we design temporal decoupling to improve end-to-end OCF by temporally associating instances via predicted flows. We develop an efficient multi-head network EfficientOCF to achieve 3D OCF with our devised spatiotemporally decoupled representation. A new metric, conditional IoU (C-IoU), is also introduced to provide a robust 3D OCF performance assessment, especially in datasets with missing or incomplete annotations. The experimental results demonstrate that EfficientOCF surpasses existing baseline methods on accuracy and efficiency, achieving state-of-the-art performance with a fast inference time of 82.33ms with a single GPU. Our code is released at: https://github.com/BIT-XJY/EfficientOCF.
****

DiffusionDrive: Truncated Diffusion Model for End-to-End Autonomous Driving
Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, Xinggang Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12037-12047
Recently, the diffusion model has emerged as a powerful generative technique for robotic policy learning, capable of modeling multi-mode action distributions. Leveraging its capability for end-to-end autonomous driving is a promising direction. However, the numerous denoising steps in the robotic diffusion policy and the more dynamic, open-world nature of traffic scenes pose substantial challenges for generating diverse driving actions at a real-time speed. To address these challenges, we propose a novel truncated diffusion policy that incorporates prior multi-mode anchors and truncates the diffusion schedule, enabling the model to learn denoising from anchored Gaussian distribution to the multi-mode driving action distribution. Additionally, we design an efficient cascade diffusion decoder for enhanced interaction with conditional scene context. The proposed model, DiffusionDrive, demonstrates 10x reduction in denoising steps compared to vanilla diffusion policy, delivering superior diversity and quality in just 2 steps. On the planning-oriented NAVSIM dataset, with the aligned ResNet-34 backbone, DiffusionDrive achieves 88.1 PDMS without bells and whistles, setting a new record, while running at a real-time speed of 45 FPS on an NVIDIA 4090. Qualitative results on challenging scenarios further confirm that DiffusionDrive can robustly generate diverse plausible driving actions.
****

MAC-Ego3D: Multi-Agent Gaussian Consensus for Real-Time Collaborative Ego-Motion and Photorealistic 3D Reconstruction
Xiaohao Xu, Feng Xue, Shibo Zhao, Yike Pan, Sebastian Scherer, Xiaonan Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 854-863
Real-time multi-agent collaboration for ego-motion estimation and high-fidelity 3D reconstruction is vital for scalable spatial intelligence. However, traditional methods produce sparse, low-detail maps, while recent dense mapping approaches struggle with high latency.To overcome these challenges, we present MAC-Ego3D, a novel framework for real-time collaborative photorealistic 3D reconstruction via Multi-Agent Gaussian Consensus. MAC-Ego3D enables agents to independently construct, align, and iteratively refine local maps using a unified Gaussian splat representation. Through Intra-Agent Gaussian Consensus, it enforces spatial coh

erence among neighboring Gaussian splats within an agent. For global alignment, parallelized Inter-Agent Gaussian Consensus, which asynchronously aligns and optimizes local maps by regularizing multi-agent Gaussian splats, seamlessly integrates them into a high-fidelity 3D model. Leveraging Gaussian primitives, MAC-Ego3D supports efficient RGB-D rendering, enabling rapid inter-agent Gaussian association and alignment.MAC-Ego3D bridges local precision and global coherence, delivering higher efficiency, largely reducing localization error, and improving mapping fidelity. It establishes a new SOTA on synthetic and real-world benchmarks, achieving a 15xincrease in inference speed, order-of-magnitude reductions in ego-motion estimation error for partial cases, and RGB PSNR gains of 4 to 10 dB. Our code will be made publicly available.

****

FFR: Frequency Feature Rectification for Weakly Supervised Semantic Segmentation
Ziqian Yang, Xinqiao Zhao, Xiaolei Wang, Quan Zhang, Jimin Xiao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30261-30270

Image-level Weakly Supervised Semantic Segmentation (WSSS) has garnered significant attention due to its low annotation costs. Current single-stage state-of-the-art WSSS methods mainly rely on V ision T ransformer (ViT) to extract features from input images, generating more complete segmentation results based on comprehensive semantic information. However, these ViT-based methods often suffer from over-smoothing issues in segmentation results. In this paper, we identify that attenuated high-frequency features mislead the decoder of ViT-based WSSS models, resulting in over-smoothed false segmentation. To address this, we propose a Frequency Feature Rectification (FFR) framework to rectify the false segmentations caused by attenuated high-frequency features and enhance the learning of high-frequency features in the decoder. Quantitative and qualitative experimental results demonstrate that our FFR framework can effectively address the attenuated high-frequency caused over-smoothed segmentation issue and achieve new state-of-the-art WSSS performances. Codes are available at https://github.com/yay97/FFR.

****

DVHGNN: Multi-Scale Dilated Vision HGNN for Efficient Vision Recognition
Caoshuo Li, Tanzhe Li, Xiaobin Hu, Donghao Luo, Taisong Jin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20158-20168

Recently, Vision Graph Neural Network (ViG) has gained considerable attention in computer vision. Despite its groundbreaking innovation, Vision Graph Neural Network encounters key issues including the quadratic computational complexity caused by its K-Nearest Neighbor (KNN) graph construction and the limitation of pairwise relations of normal graphs. To address the aforementioned challenges, we propose a novel vision architecture, termed Dilated Vision HyperGraph Neural Network (DVHGNN), which is designed to leverage multi-scale hypergraph to efficiently capture high-order correlations among objects. Specifically, the proposed method tailors Clustering and Dilated HyperGraph Construction (DHGC) to adaptively capture multi-scale dependencies among the data samples. Furthermore, a dynamic hypergraph convolution mechanism is proposed to facilitate adaptive feature exchange and fusion at the hypergraph level. Extensive qualitative and quantitative evaluations of the benchmark image datasets demonstrate that the proposed DVHGNN significantly outperforms the state-of-the-art vision backbones. For instance, our DVHGNN-S achieves an impressive top-1 accuracy of 83.1% on ImageNet-1K, surpassing ViG-S by +1.0% and ViHGNN-S by +0.6%.

****

Video-XL: Extra-Long Vision Language Model for Hour-Scale Video Understanding
Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, Bo Zhao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26160-26169

Long video understanding poses a significant challenge for current Multi-modal Large Language Models (MLLMs). Notably, the MLLMs are constrained by their limited context lengths and the substantial costs while processing long videos. Although several existing methods attempt to reduce visual tokens, their strategies encounter severe bottleneck, restricting MLLMs' ability to perceive fine-grained v

isual details. In this work, we propose Video-XL, a novel approach that leverage s MLLMs' inherent key-value (KV) sparsification capacity to condense the visual input. Specifically, we introduce a new special token, the Visual Summarization Token (VST), for each interval of the video, which summarizes the visual informa tion within the interval as its associated KV. The VST module is trained by inst ruction fine-tuning, where two optimizing strategies are offered. 1. Curriculum learning, where VST learns to make small (easy) and large compression (hard) pro gressively. 2. Composite data curation, which integrates single-image, multi-ima ge, and synthetic data to overcome the scarcity of long-video instruction data. The compression quality is further improved by dynamic compression, which custom izes compression granularity based on the information density of different video intervals. Video-XL's effectiveness is verified from three aspects. First, it a chieves a superior long-video understanding capability, outperforming state-of-t he-art models of comparable sizes across multiple popular benchmarks. Second, it effectively preserves video information, with minimal compression loss even at 16x compression ratio. Third, it realizes outstanding cost-effectiveness, enabli ng high-quality processing of thousands of frames on a single A100 GPU.
****

Reconstructing In-the-Wild Open-Vocabulary Human-Object Interactions
Boran Wen, Dingbang Huang, Zichen Zhang, Jiahong Zhou, Jianbin Deng, Jingyu Gong , Yulong Chen, Lizhuang Ma, Yong-Lu Li; Proceedings of the Computer Vision and P attern Recognition Conference (CVPR), 2025, pp. 17426-17436
Reconstructing human-object interactions (HOI) from single images is fundamental in computer vision. Existing methods are primarily trained and tested on indoor scenes due to the lack of 3D data, particularly constrained by the object varie ty, making it challenging to generalize to real-world scenes with a wide range o f objects. The limitations of previous 3D HOI datasets were primarily due to the difficulty in acquiring 3D object assets. However, with the development of 3D r econstruction from single images, recently it has become possible to reconstruct various objects from 2D HOI images.We therefore propose a pipeline for annotati ng fine-grained 3D humans, objects, and their interactions from single images. W e annotated 2.5k+ 3D HOI assets from existing 2D HOI datasets and built the firs t open-vocabulary in-the-wild 3D HOI dataset Open3DHOI, to serve as a future tes t set. Moreover, we design a novel Gaussian-HOI optimizer, which efficiently rec onstructs the spatial interactions between humans and objects while learning the contact regions.Besides the 3D HOI reconstruction, we also propose several new tasks for 3D HOI understanding to pave the way for future work.
****

GROVE: A Generalized Reward for Learning Open-Vocabulary Physical Skill
Jieming Cui, Tengyu Liu, Ziyu Meng, Jiale Yu, Ran Song, Wei Zhang, Yixin Zhu, Si yuan Huang; Proceedings of the Computer Vision and Pattern Recognition Conferenc e (CVPR), 2025, pp. 15781-15790
Learning open-vocabulary physical skills for simulated agents presents a signifi cant challenge in artificial intelligence. Current reinforcement learning approa ches face critical limitations: manually designed rewards lack scalability acros s diverse tasks, while demonstration-based methods struggle to generalize beyond their training distribution. We introduce GROVE, a generalized reward framework that enables open-vocabulary physical skill learning without manual engineering or task-specific demonstrations. Our key insight is that Large Language Models (LLMs) and Vision Language Models (VLMs) provide complementary guidance---LLMs g enerate precise physical constraints capturing task requirements, while VLMs eva luate motion semantics and naturalness. Through an iterative design process, VLM -based feedback continuously refines LLM-generated constraints, creating a self-improving reward system. To bridge the domain gap between simulation and natural images, we develop Pose2CLIP, a lightweight mapper that efficiently projects ag ent poses directly into semantic feature space without computationally expensive rendering. Extensive experiments across diverse embodiments and learning paradi gms demonstrate GROVE's effectiveness, achieving 22.2% higher motion naturalness and 25.7% better task completion scores while training 8.4x faster than previou s methods. These results establish a new foundation for scalable physical skill

acquisition in simulated environments.
****
Sonata: Self-Supervised Learning of Reliable Point Representations
Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, Julian Straub; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22193-22204
In this paper, we question whether we have a reliable self-supervised point cloud model that can be used for diverse 3D tasks via simple linear probing, even with limited data and minimal computation. We find that existing 3D self-supervised learning approaches fall short when evaluated on representation quality through linear probing. We hypothesize that this is due to what we term the "geometric shortcut", which causes representations to collapse to low-level spatial features. This challenge is unique to 3D and arises from the sparse nature of point cloud data. We address it through two key strategies: obscuring spatial information and enhancing the reliance on input features, ultimately composing a Sonata of 140k point clouds through self-distillation. Sonata is simple and intuitive, yet its learned representations are strong and reliable: zero-shot visualizations demonstrate semantic grouping, alongside strong spatial reasoning through nearest-neighbor relationships. Sonata demonstrates exceptional parameter and data efficiency, tripling linear probing accuracy (from 21.8% to 72.5%) on ScanNet and nearly doubling performance with only 1% of the data compared to previous approaches. Full fine-tuning further advances SOTA across both 3D indoor and outdoor perception tasks.
****
DriveGEN: Generalized and Robust 3D Detection in Driving via Controllable Text-to-Image Diffusion Generation
Hongbin Lin, Zilu Guo, Yifan Zhang, Shuaicheng Niu, Yafeng Li, Ruimao Zhang, Shuguang Cui, Zhen Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27497-27507
In autonomous driving, vision-centric 3D detection aims to identify 3D objects from images. However, high data collection costs and diverse real-world scenarios limit the scale of training data. Once distribution shifts occur between training and test data, existing methods often suffer from performance degradation, known as Out-of-Distribution (OOD) problems. To address this, controllable Text-to-Image (T2I) diffusion offers a potential solution for training data enhancement, which is required to generate diverse OOD scenarios with precise 3D object geometry. Nevertheless, existing controllable T2I approaches are restricted by the limited scale of training data or struggle to preserve all annotated 3D objects. In this paper, we present DriveGEN, a method designed to improve the robustness of 3D detectors in Driving via Training-Free Controllable Text-to-Image Diffusion Generation. Without extra diffusion model training, DriveGEN consistently preserves objects with precise 3D geometry across diverse OOD generations, consisting of 2 stages: 1) Self-Prototype Extraction: We empirically find that self-attention features are semantic-aware but require accurate region selection for 3D objects. Thus, we extract precise object features via layouts to capture 3D object geometry, termed self-prototypes. 2) Prototype-Guided Diffusion: To preserve objects across various OOD scenarios, we perform semantic-aware feature alignment and shallow feature alignment during denoising. Extensive experiments demonstrate the effectiveness of DriveGEN in improving 3D detection. The code is available at github.com/Hongbin98/DriveGEN.
****
GauSTAR: Gaussian Surface Tracking and Reconstruction
Chengwei Zheng, Lixin Xue, Juan Zarate, Jie Song; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16543-16553
3D Gaussian Splatting techniques have enabled efficient photo-realistic rendering of static scenes. Recent works have extended these approaches to support surface reconstruction and tracking. However, tracking dynamic surfaces with 3D Gaussians remains challenging due to complex topology changes, such as surfaces appearing, disappearing, or splitting. To address these challenges, we propose GauSTAR, a novel method that achieves photo-realistic rendering, accurate surface reco

nstruction, and reliable 3D tracking for general dynamic scenes with changing to pology. Given multi-view captures as input, GauSTAR binds Gaussians to mesh face s to represent dynamic objects. For surfaces with consistent topology, GauSTAR m aintains the mesh topology and tracks the meshes using Gaussians. For regions wh ere topology changes, GauSTAR adaptively unbinds Gaussians from the mesh, enabli ng accurate registration and generation of new surfaces based on these optimized Gaussians. Additionally, we introduce a surface-based scene flow method that pr ovides robust initialization for tracking between frames. Experiments demonstrat e that our method effectively tracks and reconstructs dynamic surfaces, enabling a range of applications. Our project page with the code release is available at https://eth-ait.github.io/GauSTAR/.

****

Training-free Dense-Aligned Diffusion Guidance for Modular Conditional Image Syn thesis

Zixuan Wang, Duo Peng, Feng Chen, Yuwei Yang, Yinjie Lei; Proceedings of the Com puter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13135-13145

Conditional image synthesis is a crucial task with broad applications, such as a rtistic creation and virtual reality. However, current generative methods are of ten task-oriented with a narrow scope, handling a restricted condition with cons trained applicability. In this paper, we propose a novel approach that treats co nditional image synthesis as the modular combination of diverse fundamental cond ition units. Specifically, we divide conditions into three primary units: text, layout, and drag. To enable effective control over these conditions, we design a dedicated alignment module for each. For the text condition, we introduce a Den se Concept Alignment (DCA) module, which achieves dense visual-text alignment by drawing on diverse textual concepts. For the layout condition, we propose a Den se Geometry Alignment (DGA) module to enforce comprehensive geometric constraint s that preserve the spatial configuration. For the drag condition, we introduce a Dense Motion Alignment (DMA) module to apply multi-level motion regularization , ensuring that each pixel follows its desired trajectory without visual artifac ts. By flexibly inserting and combining these alignment modules, our framework e nhances the model's adaptability to diverse conditional generation tasks and gre atly expands its application range. Extensive experiments demonstrate the superi or performance of our framework across a variety of conditions, including textua l description, segmentation mask (bounding box), drag manipulation, and their co mbinations. Code is available at https://github.com/ZixuanWang0525/DADG

****

DRiVE: Diffusion-based Rigging Empowers Generation of Versatile and Expressive C haracters

Mingze Sun, Junhao Chen, Junting Dong, Yurun Chen, Xinyu Jiang, Shiwei Mao, Puhu a Jiang, Jingbo Wang, Bo Dai, Ruqi Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21170-21180

Recent advances in generative models have enabled high-quality 3D character reco nstruction from multi-modal. However, animating these generated characters remai ns a challenging task, especially for complex elements like garments and hair, d ue to the lack of large-scale datasets and effective rigging methods. To address this gap, we curate AnimeRig, a large-scale dataset with detailed skeleton and skinning annotations. Building upon this, we propose DRiVE, a novel framework fo r generating and rigging 3D human characters with intricate structures. Unlike e xisting methods, DRiVE utilizes a 3D Gaussian representation, facilitating effic ient animation and high-quality rendering. We further introduce GSDiff, a 3D Gau ssian-based diffusion module that predicts joint positions as spatial distributi ons, overcoming the limitations of regression-based approaches. Extensive experi ments demonstrate that DRiVE achieves precise rigging results, enabling realisti c dynamics for clothing and hair, and surpassing previous methods in both qualit y and versatility. Code, dataset and visualization results are available at http s://DRiVEAvatar.github.io/.

****

Online Video Understanding: OVBench and VideoChat-Online

Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao W

u, Xi Chen, Liang Li, Limin Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3328-3338
Multimodal Large Language Models (MLLMs) have significantly progressed in offline video understanding. However, applying these models to real-world scenarios, such as autonomous driving and human-computer interaction, presents unique challenges due to the need for real-time processing of continuous online video streams. To this end, this paper presents systematic efforts from three perspectives: evaluation benchmark, model architecture, and training strategy. First, we introduce OVBench, a comprehensive question-answering benchmark designed to evaluate models' ability to perceive, memorize, and reason within online video contexts. It features 6 core task types across three temporal contexts--past, current, and future--forming 16 subtasks from diverse datasets. Second, we propose a new Pyramid Memory Bank (PMB) that effectively retains key spatiotemporal information in video streams. Third, we proposed an offline-to-online learning paradigm, designing an interleaved dialogue format for online video data and constructing an instruction-tuning dataset tailored for online video training. This framework led to the development of VideoChat-Online, a robust and efficient model for online video understanding. Despite the lower computational cost and higher efficiency, VideoChat-Online outperforms existing state-of-the-art offline and online models across popular offline video benchmarks and OVBench, demonstrating the effectiveness of our model architecture and training strategy.
****

TADFormer: Task-Adaptive Dynamic TransFormer for Efficient Multi-Task Learning
Seungmin Baek, Soyul Lee, Hayeon Jo, Hyesong Choi, Dongbo Min; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14858-14868
Transfer learning paradigm has driven substantial advancements in various vision tasks. However, as state-of-the-art models continue to grow, classical full fine-tuning often becomes computationally impractical, particularly in multi-task learning (MTL) setup where training complexity increases proportional to the number of tasks. Consequently, recent studies have explored Parameter-Efficient Fine-Tuning (PEFT) for MTL architectures. Despite some progress, these approaches still exhibit limitations in capturing fine-grained, task-specific features that are crucial to MTL. In this paper, we introduce Task-Adaptive Dynamic transFormer, termed TADFormer, a novel PEFT framework that performs task-aware feature adaptation in the fine-grained manner by dynamically considering task-specific input contexts. TADFormer proposes the parameter-efficient prompting for task adaptation and the Dynamic Task Filter (DTF) to capture task information conditioned on input contexts. Experiments on the PASCAL-Context benchmark demonstrate that the proposed method achieves higher accuracy in dense scene understanding tasks, while reducing the number of trainable parameters by up to 8.4 times when compared to full fine-tuning of MTL models. TADFormer also demonstrates superior parameter efficiency and accuracy compared to recent PEFT methods. Our code is available at supplementary material.
****

A Unified Approach to Interpreting Self-supervised Pre-training Methods for 3D Point Clouds via Interactions
Qiang Li, Jian Ruan, Fanghao Wu, Yuchi Chen, Zhihua Wei, Wen Shen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27315-27324
Recently, many self-supervised pre-training methods have been proposed to improve the performance of deep neural networks (DNNs) for 3D point clouds processing. However, the common mechanism underlying the effectiveness of different pre-training methods remains unclear. In this paper, we use game-theoretic interactions as a unified approach to explore the common mechanism of pre-training methods. Specifically, we decompose the output score of a DNN into the sum of numerous effects of interactions, with each interaction representing a distinct 3D substructure of the input point cloud. Based on the decomposed interactions, we draw the following conclusions. (1) The common mechanism across different pre-training methods is that they enhance the strength of high-order interactions encoded by D

NNs, which represent complex and global 3D structures, while reducing the streng th of low-order interactions, which represent simple and local 3D structures. (2) Sufficient pre-training and adequate fine-tuning data for downstream tasks fur ther reinforce the mechanism described above. (3) Pre-training methods carry a p otential risk of reducing the transferability of features encoded by DNNs. Inspi red by the observed common mechanism, we propose a new method to directly enhanc e the strength of high-order interactions and reduce the strength of low-order i nteractions encoded by DNNs, improving performance without the need for pre-trai ning on large-scale datasets. Experiments show that our method achieves performa nce comparable to traditional pre-training methods.
****

Enhancing SAM with Efficient Prompting and Preference Optimization for Semi-supe rvised Medical Image Segmentation
Aishik Konwer, Zhijian Yang, Erhan Bas, Cao Xiao, Prateek Prasanna, Parminder Bh atia, Taha Kass-Hout; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20990-21000
Foundational models such as the Segment Anything Model (SAM) are gaining tractio n in medical imaging segmentation, supporting multiple downstream tasks. However , such models are supervised in nature, still relying on large annotated dataset s or prompts supplied by experts. Conventional techniques such as active learnin g to alleviate such limitations are limited in scope and still necessitate conti nuous human involvement and complex domain knowledge for label refinement or est ablishing reward ground truth. To address these challenges, we propose an enhanc ed Segment Anything Model (SAM) framework that utilizes annotation-efficient pro mpts generated in a fully unsupervised fashion, while still capturing essential semantic, location, and shape information through contrastive language-image pre training and visual question answering. We adopt the direct preference optimizat ion technique to design an optimal policy that enables the model to generate hig h-fidelity segmentations with simple ratings or rankings provided by a virtual a nnotator simulating the human annotation process. State-of-the-art performance o f our framework in tasks such as lung segmentation, breast tumor segmentation, a nd organ segmentation across various modalities, including X-ray, ultrasound, an d abdominal CT, justifies its effectiveness in low-annotation data scenarios.
****

CamPoint: Boosting Point Cloud Segmentation with Virtual Camera
Jianhui Zhang, Yizhi Luo, Zicheng Zhang, Xuecheng Nie, Bonan Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11822-1 1832
Local features aggregation and global information perception are the fundamental to point cloud segmentation. However, existing works often fall short in effect ively identifying semantic relevant neighbors and face challenges in endowing ea ch point with high-level information. Here, we propose CamPoint, an innovative m ethod that employs virtual cameras to solve above problems. The core of CamPoint lies in introducing the novel camera visibility feature for points, where each dimension encodes the visibility of that point from a specific camera. Leveragin g this feature, we propose the camera perspective slice distance for accurate re levant neighbor searching and design the camera parameter embedding to deliver r ich feature representations for global interaction. Specifically, the camera per spective slice distance between two points is defined as a similarity metric der ived from their camera visibility features, whereby an increased number of share d cameras observing both points corresponds to a reduced distance between them. To effectively facilitate global semantic perception, we assign each camera an o ptimizable embedding and then integrate these embeddings into the original spati al features based on visibility attributes, thereby obtaining high-level feature s enriched with camera priors. Additionally, the state space model characterized by linear computational complexity is employed as the operator to achieve globa l learning with efficiency. Comprehensive experiments on multiple datasets shows that our CamPoint surpasses the current state-of-the-art in multiple datasets, achieving low training cost and fast inference speed.
****

LightLoc: Learning Outdoor LiDAR Localization at Light Speed
Wen Li, Chen Liu, Shangshu Yu, Dunqiang Liu, Yin Zhou, Siqi Shen, Chenglu Wen, Cheng Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6680-6689
Scene coordinate regression achieves impressive results in outdoor LiDAR localization but requires days of training. Since training needs to be repeated for each new scene, long training times make these impractical for applications requiring time-sensitive system upgrades, such as autonomous driving, drones, robotics, etc. We identify large coverage areas and vast amounts of data in large-scale outdoor scenes as key challenges that limit fast training. In this paper, we propose LightLoc, the first method capable of efficiently learning localization in a new scene at light speed. Beyond freezing the scene-agnostic feature backbone and training only the scene-specific prediction heads, we introduce two novel techniques to address these challenges. First, we introduce sample classification guidance to assist regression learning, reducing ambiguity from similar samples and improving training efficiency. Second, we propose redundant sample downsampling to remove well-learned frames during training, reducing training time without compromising accuracy. In addition, the fast training and confidence estimation characteristics of sample classification enable its integration into SLAM, effectively eliminating error accumulation. Extensive experiments on large-scale outdoor datasets demonstrate that LightLoc achieves state-of-the-art performance with just 1 hour of training--50x faster than existing methods. Our Code is available at https://github.com/liw95/LightLoc.
****
MERGE: Multi-faceted Hierarchical Graph-based GNN for Gene Expression Prediction from Whole Slide Histopathology Images
Aniruddha Ganguly, Debolina Chatterjee, Wentao Huang, Jie Zhang, Alisa Yurovsky, Travis Steele Johnson, Chao Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15611-15620
Recent advances in Spatial Transcriptomics (ST) pair histology images with spatially resolved gene expression profiles, enabling predictions of gene expression across different tissue locations based on image patches. This opens up new possibilities for enhancing whole slide image (WSI) prediction tasks with localized gene expression. However, existing methods fail to fully leverage the interactions between different tissue locations, which are crucial for accurate joint prediction. To address this, we introduce MERGE (Multi-faceted hiErarchical gRaph for Gene Expressions), which combines a multi-faceted hierarchical graph construction strategy with graph neural networks (GNN) to improve gene expression predictions from WSIs. By clustering tissue image patches based on both spatial and morphological features, and incorporating intra- and inter-cluster edges, our approach fosters interactions between distant tissue locations during GNN learning. As an additional contribution, we evaluate different data smoothing techniques that are necessary to mitigate artifacts in ST data, often caused by technical imperfections. We advocate for adopting gene-aware smoothing methods that are more biologically justified. Experimental results on gene expression prediction show that our GNN method outperforms state-of-the-art techniques across multiple metrics.
****
Accurate Differential Operators for Hybrid Neural Fields
Aditya Chetan, Guandao Yang, Zichen Wang, Steve Marschner, Bharath Hariharan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 530-539
Neural fields have become widely used in various fields, from shape representation to neural rendering, and for solving partial differential equations (PDEs). With the advent of hybrid neural field representations like Instant NGP that leverage small MLPs and explicit representations, these models train quickly and can fit large scenes. Yet in many applications like rendering and simulation, hybrid neural fields can cause noticeable and unreasonable artifacts. This is because they do not yield accurate spatial derivatives needed for these downstream applications. In this work, we propose two ways to circumvent these challenges. Our

first approach is a post hoc operator that uses local polynomial fitting to obta in more accurate derivatives from pre-trained hybrid neural fields. Additionally , we also propose a self-supervised fine-tuning approach that refines the hybrid neural field to yield accurate derivatives directly while preserving the initia l signal. We show applications of our method to rendering, collision simulation, and solving PDEs. We observe that using our approach yields more accurate deriv atives, reducing artifacts and leading to more accurate simulations in downstrea m applications.
****

FeedEdit: Text-Based Image Editing with Dynamic Feedback Regulation
Fengyi Fu, Lei Zhang, Mengqi Huang, Zhendong Mao; Proceedings of the Computer Vi sion and Pattern Recognition Conference (CVPR), 2025, pp. 2661-2670
Text-based image editing which aims at generating rigid or non-rigid changes to images conditioned on the given text, has recently attracted considerable intere st. Previous works mainly follow the multi-step denoising diffusion paradigm, wh ich adopts a fixed text guidance intensity (i.e., editing intensity) to inject t extual features, while ignoring the step-specific editing requirements. This wor k argues that the editing intensity at each denoising step should be adaptively adjusted conditioned on the historical editing degree, to provide accurate text guidance for the whole denoising process. We thereby propose a novel feedback ed iting framework (FeedEdit), a training-free method which, explicitly exploits th e feedback regulation on editing intensity to ensure precise and harmonious edit ing at all steps. Specifically, we design (1) Dynamic Editing Degree Perceiving module, which is based on specific frequency-domain filtering, to enhance and ex ploit the correlation between feature differences and editing degree for perceiv ing. (2) Proportional-Integral feedback controller, to automatically map the pe rceived editing errors into appropriate feedback control signals. (3) Phrase-lev el Regulating Strategy, to achieve fine-grained function-specific regulation of textual features. Extensive experiments demonstrate the superiority of FeedEdit over existing methods in both editability and quality, especially for multi-func tion editing scenarios.
****

Classifier-guided CLIP Distillation for Unsupervised Multi-label Classification
Dongseob Kim, Hyunjung Shim; Proceedings of the Computer Vision and Pattern Reco gnition Conference (CVPR), 2025, pp. 4661-4671
Multi-label classification is crucial for comprehensive image understanding, yet acquiring accurate annotations is challenging and costly. To address this, a re cent study suggests exploiting unsupervised multi-label classification leveragin g CLIP, a powerful vision-language model. Despite CLIP's proficiency, it suffers from view-dependent predictions and inherent bias, limiting its effectiveness. We propose a novel method that addresses these issues by leveraging multiple vie ws near target objects, guided by Class Activation Mapping (CAM) of the classifi er, and debiasing pseudo-labels derived from CLIP predictions. Our Classifier-gu ided CLIP Distillation (CCD) enables selecting multiple local views without extr a labels and debiasing predictions to enhance classification performance. Experi mental results validate our method's superiority over existing techniques across diverse datasets. The code is available at https://github.com/k0u-id/CCD.
****

UMotion: Uncertainty-driven Human Motion Estimation from Inertial and Ultra-wide band Units
Huakun Liu, Hiroki Ota, Xin Wei, Yutaro Hirao, Monica Perusquia-Hernandez, Hidea ki Uchiyama, Kiyoshi Kiyokawa; Proceedings of the Computer Vision and Pattern Re cognition Conference (CVPR), 2025, pp. 7085-7094
Sparse wearable inertial measurement units (IMUs) have gained popularity for est imating 3D human motion. However, challenges such as pose ambiguity, data drift, and limited adaptability to diverse bodies persist. To address these issues, we propose UMotion, an uncertainty-driven, online fusing-all state estimation fram ework for 3D human shape and pose estimation, supported by six integrated, body-worn ultra-wideband (UWB) distance sensors with IMUs. UWB sensors measure inter-node distances to infer spatial relationships, aiding in resolving pose ambiguit

ies and body shape variations when combined with anthropometric data. Unfortunat ely, IMUs are prone to drift, and UWB sensors are affected by body occlusions. C onsequently, we develop a tightly coupled Unscented Kalman Filter (UKF) framewor k that fuses uncertainties from sensor data and estimated human motion based on individual body shape. The UKF iteratively refines IMU and UWB measurements by a ligning them with uncertain human motion constraints in real-time, producing opt imal estimates for each. Experiments on both synthetic and real-world datasets d emonstrate the effectiveness of UMotion in stabilizing sensor data and the impro vement over state of the art in pose accuracy. Code is available at: https://git hub.com/kk9six/umotion.
****

STEREO: A Two-Stage Framework for Adversarially Robust Concept Erasing from Text -to-Image Diffusion Models
Koushik Srivatsan, Fahad Shamshad, Muzammal Naseer, Vishal M. Patel, Karthik Nan dakumar; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23765-23774
The rapid proliferation of large-scale text-to-image diffusion (T2ID) models has raised serious concerns about their potential misuse in generating harmful cont ent. Although numerous methods have been proposed for erasing undesired concept s from T2ID models, they often provide a false sense of security; concept-erased models (CEMs) can still be manipulated via adversarial attacks to regenerate th e erased concept. While a few robust concept erasure methods based on adversaria l training have emerged recently, they compromise on utility (generation quality for benign concepts) to achieve robustness and/or remain vulnerable to advanced embedding space attacks. These limitations stem from the failure of robust CEMs to thoroughly search for blind spots in the embedding space. To bridge this gap , we propose STEREO, a novel two-stage framework that employs adversarial traini ng as a first step rather than the only step for robust concept erasure. In the first stage, STEREO employs adversarial training as a vulnerability identificati on mechanism to search thoroughly enough. In the second robustly erase once stag e, STEREO introduces an anchor-concept-based compositional objective to robustly erase the target concept in a single fine-tuning stage, while minimizing the de gradation of model utility. We benchmark STEREO against seven state-of-the-art c oncept erasure methods, demonstrating its superior robustness to both white-box and black-box attacks, while largely preserving utility.
****

Scene Map-based Prompt Tuning for Navigation Instruction Generation
Sheng Fan, Rui Liu, Wenguan Wang, Yi Yang; Proceedings of the Computer Vision an d Pattern Recognition Conference (CVPR), 2025, pp. 6898-6908
Navigation instruction generation (NIG), which provides interactive feedback and guidance to humans along a trajectory, is vital for developing embodied agents capable of human-machine communication and collaboration through natural languag e. Early data-driven methods directly map sequences of past observations to traj ectory descriptions on limited datasets, lacking the necessary spatial understan ding in complex 3D environments. While recent approaches leverage Large Language Models (LLMs) to improve NIG, they often overlook the global spatial context in navigation, such as the inherent space discretization in maps. Instead of strai ghtforwardly feeding textual descriptions of the map into LLMs, we propose a sce ne map-based prompt tuning framework for NIG, MAPInstructor, which incorporates map context for parameter-efficient updating of LLMs. MAPInstructor comprises th ree key components: (i) scene representation encoding, where egocentric observat ions are projected into 3D voxels for fine-grained scene understanding; (ii) map prompt tuning, which integrates a topological map representation of the entire trajectory into an LLM-based decoder; and (iii) landmark uncertainty assessment, which mitigates hallucinations in landmark predictions, thereby enhancing the r eliability and coherence of instruction generation. Extensive experiments on thr ee navigation datasets (i.e., R2R, REVERIE, RxR) confirm the generalization and effectiveness of our algorithm.
****

GenVDM: Generating Vector Displacement Maps From a Single Image

Yuezhi Yang, Qimin Chen, Vladimir G. Kim, Siddhartha Chaudhuri, Qixing Huang, Zh
iqin Chen; Proceedings of the Computer Vision and Pattern Recognition Conference
 (CVPR), 2025, pp. 26618-26629

We introduce the first method for generating Vector Displacement Maps (VDMs): pa
rameterized, detailed geometric stamps commonly used in 3D modeling. Given a sin
gle input image, our method first generates multi-view normal maps and then reco
nstructs a VDM from the normals via a novel reconstruction pipeline. We also pro
pose an efficient algorithm for extracting VDMs from 3D objects, and present the
 first academic VDM dataset. Compared to existing 3D generative models focusing
on complete shapes, we focus on generating parts that can be seamlessly attached
 to shape surfaces. The method gives artists rich control over adding geometric
details to a 3D shape. Experiments demonstrate that our approach outperforms exi
sting baselines. Generating VDMs offers additional benefits, such as using 2D im
age editing to customize and refine 3D details.

****

DropoutGS: Dropping Out Gaussians for Better Sparse-view Rendering

Yexing Xu, Longguang Wang, Minglin Chen, Sheng Ao, Li Li, Yulan Guo; Proceedings
 of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 701
-710

Although 3D Gaussian Splatting (3DGS) has demonstrated promising results in nove
l view synthesis, its performance degrades dramatically with sparse inputs and g
enerates undesirable artifacts. As the number of training views decreases, the n
ovel view synthesis task degrades to a highly under-determined problem such that
 existing methods suffer from the notorious overfitting issue. Interestingly, we
 observe that models with fewer Gaussian primitives exhibit less overfitting und
er spare inputs. Inspired by this observation, we propose a Random Dropout Regul
arization (RDR) to exploit the advantages of low-complexity models to alleviate
overfitting. In addition, to remedy the lack of high-frequency details for these
 models, an Edge-guided Splitting Strategy (ESS) is developed. With these two te
chniques, our method (termed DropoutGS) provides a simple yet effective plug-in
approach to improve the generalization performance of existing 3DGS methods. Ext
ensive experiments show that our DropoutGS produces state-of-the-art performance
 under sparse views on benchmark datasets including Blender, LLFF, and DTU.

****

Effective SAM Combination for Open-Vocabulary Semantic Segmentation

Minhyeok Lee, Suhwan Cho, Jungho Lee, Sunghun Yang, Heeseung Choi, Ig-Jae Kim, S
angyoun Lee; Proceedings of the Computer Vision and Pattern Recognition Conferen
ce (CVPR), 2025, pp. 26081-26090

Open-vocabulary semantic segmentation aims to assign pixel-level labels to image
s across an unlimited range of classes. Traditional methods address this by sequ
entially connecting a powerful mask proposal generator, such as the Segment Anyt
hing Model (SAM), with a pre-trained vision-language model like CLIP. But these
two-stage approaches often suffer from high computational costs, memory ineffici
encies. In this paper, we propose ESC-Net, a novel one-stage open-vocabulary seg
mentation model that leverages the SAM decoder blocks for class-agnostic segment
ation within an efficient inference framework. By embedding pseudo prompts gener
ated from image-text correlations into SAM's promptable segmentation framework,
ESC-Net achieves refined spatial aggregation for accurate mask predictions. Addi
tionally, a Vision-Language Fusion (VLF) module enhances the final mask predicti
on through image and text guidance. ESC-Net achieves superior performance on sta
ndard benchmarks, including ADE20K, PASCAL-VOC, and PASCAL-Context, outperformin
g prior methods in both efficiency and accuracy. Comprehensive ablation studies
further demonstrate its robustness across challenging conditions.

****

Bridging Modalities: Improving Universal Multimodal Retrieval by Multimodal Larg
e Language Models

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun X
ie, Meishan Zhang, Wenjie Li, Min Zhang; Proceedings of the Computer Vision and
Pattern Recognition Conference (CVPR), 2025, pp. 9274-9285

Universal Multimodal Retrieval (UMR) aims to enable search across various modali

ties using a unified model, where queries and candidates can consist of pure tex
t, images, or a combination of both. Previous work has attempted to adopt multim
odal large language models (MLLMs) to realize UMR using only text data. However,
 our preliminary experiments demonstrate that more diverse multimodal training d
ata can further unlock the potential of MLLMs. Despite its effectiveness, the ex
isting multimodal training data is highly imbalanced in terms of modality, which
 motivates us to develop a training data synthesis pipeline and construct a larg
e-scale, high-quality fused-modal training dataset. Based on the synthetic train
ing data, we develop the General Multimodal Embedder (GME), an MLLM-based dense
retriever designed for UMR. Furthermore, we construct a comprehensive UMR Benchm
ark (UMRB) to evaluate the effectiveness of our approach. Experimental results s
how that our method achieves state-of-the-art performance among existing UMR met
hods. Last, we provide in-depth analyses of model scaling and training strategie
s, and perform ablation studies on both the model and synthetic data.
****

Enhancing Dataset Distillation via Non-Critical Region Refinement
Minh-Tuan Tran, Trung Le, Xuan-May Le, Thanh-Toan Do, Dinh Phung; Proceedings of
 the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10015-
10024
Dataset distillation has gained popularity as a technique for compressing large
datasets into smaller, more efficient representations while retaining essential
information for model training. Data features can be broadly divided into two t
ypes: instance-specific features, which capture unique, fine-grained details of
individual examples, and class-general features, which represent shared, broad p
atterns across a class. However, previous approaches often struggle to balance t
hese; some focus solely on class-general features, missing finer instance detail
s, while others concentrate on instance-specific features, overlooking the share
d characteristics essential for class-level understanding. In this paper, we pro
pose the Non-Critical Region Refinement Dataset Distillation (NRR-DD) method, wh
ich preserves the instance-specific and fine-grained regions in synthetic data w
hile enriching non-critical regions with more class-general information. This ap
proach enables our models to leverage all pixel information to capture both type
s of features, thereby improving overall performance. Furthermore, we introduce
Distance-Based Representative (DBR) knowledge transfer, which eliminates the nee
d for soft labels in training by relying solely on the distance between syntheti
c data predictions and one-hot encoded labels. Experimental results demonstrate
that our NRR-DD achieves state-of-the-art performance on both small-scale and la
rge-scale datasets. Additionally, by storing only two distances per instance, ou
r method achieves comparable results across various settings. Code will be avail
able at https://github.com/tmtuan1307/NRR-DD.
****

Towards Visual Discrimination and Reasoning of Real-World Physical Dynamics: Phy
sics-Grounded Anomaly Detection
Wenqiao Li, Yao Gu, Xintao Chen, Xiaohao Xu, Ming Hu, Xiaonan Huang, Yingna Wu;
Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 20
25, pp. 30409-30419
Humans detect real-world object anomalies by perceiving, interacting, and reason
ing based on object-conditioned physical knowledge. The long-term goal of Indust
rial Anomaly Detection (IAD) is to enable machines to autonomously replicate thi
s skill. However, current IAD algorithms are largely developed and tested on sta
tic, semantically simple datasets, which diverge from real-world scenarios where
 physical understanding and reasoning are essential.To bridge this gap, we intro
duce the Physics Anomaly Detection (Phys-AD) dataset, the first large-scale, rea
l-world, physics-grounded video dataset for industrial anomaly detection. Collec
ted using a real robot arm and motor, Phys-AD provides a diverse set of dynamic,
 semantically rich scenarios. The dataset includes more than 6400 videos across
22 real-world object categories, interacting with robot arms and motors, and exh
ibits 47 types of anomalies. Anomaly detection in Phys-AD requires visual reason
ing, combining both physical knowledge and video content to determine object abn
ormality.We benchmark state-of-the-art anomaly detection methods under three set

tings: unsupervised AD, weakly-supervised AD, and video-understanding AD, highli ghting their limitations in handling physics-grounded anomalies. Additionally, w e introduce the Physics Anomaly Explanation (PAEval) metric, designed to assess the ability of visual-language foundation models to not only detect anomalies bu t also provide accurate explanations for their underlying physical causes. Our p roject is available at https://guyao2023.github.io/Phys-AD/.

****

SPA-VL: A Comprehensive Safety Preference Alignment Dataset for Vision Language Models

Yongting Zhang, Lu Chen, Guodong Zheng, Yifeng Gao, Rui Zheng, Jinlan Fu, Zhenfe i Yin, Senjie Jin, Yu Qiao, Xuanjing Huang, Feng Zhao, Tao Gui, Jing Shao; Proce edings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, p p. 19867-19878

The emergence of Vision Language Models (VLMs) has brought unprecedented advance s in understanding multimodal information. The combination of textual and visual semantics in VLMs is highly complex and diverse, making the safety alignment of these models challenging. Furthermore, due to the limited study on the safety a lignment of VLMs, there is a lack of large-scale, high-quality datasets. To addr ess these limitations, we propose a Safety Preference Alignment dataset for Visi on Language Models named SPA-VL. In terms of breadth, SPA-VL covers 6 harmfulnes s domains, 13 categories, and 53 subcategories, and contains 100,788 samples of the quadruple (question, image, chosen response, rejected response). In terms of depth, the responses are collected from 12 open-source (e.g., QwenVL) and close d-source (e.g., Gemini) VLMs to ensure diversity. The construction of preference data is fully automated, and the experimental results indicate that models trai ned with alignment techniques on the SPA-VL dataset exhibit substantial improvem ents in harmlessness and helpfulness while maintaining core capabilities. SPA-VL , as a large-scale, high-quality, and diverse dataset, represents a significant milestone in ensuring that VLMs achieve both harmlessness and helpfulness.

****

PUP 3D-GS: Principled Uncertainty Pruning for 3D Gaussian Splatting

Alex Hanson, Allen Tu, Vasu Singla, Mayuka Jayawardhana, Matthias Zwicker, Tom G oldstein; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5949-5958

Recent advances in novel view synthesis have enabled real-time rendering speeds with high reconstruction accuracy. 3D Gaussian Splatting (3D-GS), a foundational point-based parametric 3D scene representation, models scenes as large sets of 3D Gaussians. However, complex scenes can consist of millions of Gaussians, resu lting in high storage and memory requirements that limit the viability of 3D-GS on devices with limited resources. Current techniques for compressing these pret rained models by pruning Gaussians rely on combining heuristics to determine whi ch Gaussians to remove. At high compression ratios, these pruned scenes suffer f rom heavy degradation of visual fidelity and loss of foreground details. In this paper, we propose a principled sensitivity pruning score that preserves visual fidelity and foreground details at significantly higher compression ratios than existing approaches. It is computed as a second-order approximation of the recon struction error on the training views with respect to the spatial parameters of each Gaussian. Additionally, we propose a multi-round prune-refine pipeline that can be applied to any pretrained 3D-GS model without changing its training pipe line. After pruning 90% of Gaussians, a substantially higher percentage than pre vious methods, our PUP 3D-GS pipeline increases average rendering speed by 3.56x while retaining more salient foreground information and achieving higher image quality metrics than existing techniques on scenes from the Mip-NeRF 360, Tanks & Temples, and Deep Blending datasets.

****

UniAP: Unifying Inter- and Intra-Layer Automatic Parallelism by Mixed Integer Qu adratic Programming

Hao Lin, Ke Wu, Jie Li, Jun Li, Wu-Jun Li; Proceedings of the Computer Vision an d Pattern Recognition Conference (CVPR), 2025, pp. 20947-20957

Distributed learning is commonly used for training deep learning models, especia

lly large models. In distributed learning, manual parallelism (MP) methods demand considerable human effort and have limited flexibility. Hence, automatic parallelism (AP) methods have recently been proposed for automating the parallel strategy optimization process. Existing AP methods suffer from sub-optimal solutions because they do not jointly optimize the two categories of parallel strategies (i.e., inter-layer parallelism and intra-layer parallelism). In this paper, we propose a novel AP method called UniAP, which unifies inter- and intra-layer automatic parallelism by mixed integer quadratic programming. To the best of our knowledge, UniAP is the first parallel method that can jointly optimize the two categories of parallel strategies to find an optimal solution. Experimental results show that UniAP outperforms state-of-the-art methods by up to 3.80x in throughput and reduces strategy optimization time by up to 107x across five Transformer-based models.

****

PTDiffusion: Free Lunch for Generating Optical Illusion Hidden Pictures with Phase-Transferred Diffusion Model

Xiang Gao, Shuai Yang, Jiaying Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18240-18249

Optical illusion hidden picture is an interesting visual perceptual phenomenon where an image is cleverly integrated into another picture. Established on the off-the-shelf text-to-image (T2I) diffusion model, we propose a novel text-guided image-to-image (I2I) translation framework dubbed as Phase-Transferred Diffusion Model (PTDiffusion) for hidden art syntheses, which harmoniously embeds an input reference image into arbitrary scenes described by the text prompts. At the heart of our method is a plug-and-play phase transfer mechanism that dynamically and progressively transplants diffusion features' phase spectrum from the denoising process to reconstruct the reference image into the one to sample the generated illusion image, realizing deep fusion of the reference structural information and the textual semantic information. Furthermore, we propose asynchronous phase transfer to enable flexible control over the degree of hidden content discernability. Our method bypasses any model training and fine-tuning process, all while substantially outperforming related methods in image quality, text fidelity, visual discernibility, and contextual naturalness for illusion picture synthesis, as demonstrated by extensive qualitative and quantitative experiments. Our project is publically available at \href https://xianggao1102.github.io/PTDiffusion_webpage/ this web page .

****

ScribbleLight: Single Image Indoor Relighting with Scribbles

Jun Myeong Choi, Annie Wang, Pieter Peers, Anand Bhattad, Roni Sengupta; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5720-5731

Image-based relighting of indoor rooms creates an immersive virtual understanding of the space, which is useful for interior design, virtual staging, and real estate. Relighting indoor rooms from a single image is especially challenging due to complex illumination interactions between multiple lights and cluttered objects featuring a large variety in geometrical and material complexity. Recently, generative models have been successfully applied to image-based relighting conditioned on a target image or a latent code, albeit without detailed local lighting control. In this paper, we introduce ScribbleLight, a generative model that supports local fine-grained control of lighting effects through scribbles that describe changes in lighting. Our key technical novelty is an Albedo-conditioned Stable Image Diffusion model that preserves the intrinsic color and texture of the original image after relighting and an encoder-decoder-based ControlNet architecture that enables geometry-preserving lighting effects with normal map and scribble annotations. We demonstrate ScribbleLight's ability to create different lighting effects (e.g., turning lights on/off, adding highlights, cast shadows, or indirect lighting from unseen lights) from sparse scribble annotations.

****

Preserving Clusters in Prompt Learning for Unsupervised Domain Adaptation

Tung-Long Vuong, Hoang Phan, Vy Vo, Anh Bui, Thanh-Toan Do, Trung Le, Dinh Phung

; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19974-19984
Recent approaches leveraging multi-modal pre-trained models like CLIP for Unsupervised Domain Adaptation (UDA) have shown significant promise in bridging domain gaps and improving generalization by utilizing rich semantic knowledge and robust visual representations learned through extensive pre-training on diverse image-text datasets. While these methods achieve state-of-the-art performance across benchmarks, much of the improvement stems from base pseudo-labels (CLIP zero-shot predictions) and self-training mechanisms. Thus, the training mechanism exhibits a key limitation wherein the visual embedding distribution in target domains can deviate from the visual embedding distribution in the pre-trained model, leading to misguided signals from class descriptions. This work introduces a fresh solution to reinforce these pseudo-labels and facilitate target-prompt learning, by exploiting the geometry of visual and text embeddings - an aspect that is overlooked by existing methods. We first propose to directly leverage the reference predictions (from source prompts) based on the relationship between source and target visual embeddings. We later show that there is a strong clustering behavior observed between visual and text embeddings in pre-trained multi-modal models. Building on optimal transport theory, we transform this insight into a novel strategy to enforce the clustering property in text embeddings, further enhancing the alignment in the target domain. Our experiments and ablation studies validate the effectiveness of the proposed approach, demonstrating superior performance and improved quality of target prompts in terms of representation.
****

InsightEdit: Towards Better Instruction Following for Image Editing
Yingjing Xu, Jie Kong, Jiazhi Wang, Xiao Pan, Bo Lin, Qiang Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2694-2703
In this paper, we focus on the task of instruction-based image editing. Previous works like InstructPix2Pix, InstructDiffusion, and SmartEdit have explored end-to-end editing. However, two limitations still remain: First, existing datasets suffer from low resolution, poor background consistency, and overly simplistic instructions. Second, current approaches mainly condition on the text while the rich image information is underexplored, therefore inferior in complex instruction following and maintaining background consistency. Targeting these issues, we first curated the AdvancedEdit dataset using a novel data construction pipeline, formulating a large-scale dataset with high visual quality, complex instructions, and good background consistency. Then, to further inject the rich image information, we introduce a two-stream bridging mechanism utilizing both the textual and visual features reasoned by the powerful Multimodal Large Language Models (MLLM) to guide the image editing process more precisely. Extensive results demonstrate that our approach, InsightEdit, achieves state-of-the-art performance, excelling in complex instruction following and maintaining high background consistency with the original image.
****

Attend to Not Attended: Structure-then-Detail Token Merging for Post-training DiT Acceleration
Haipeng Fang, Sheng Tang, Juan Cao, Enshuo Zhang, Fan Tang, Tong-Yee Lee; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18083-18092
Diffusion transformers have shown exceptional performance in visual generation but incur high computational costs. Token reduction techniques that compress models by sharing the denoising process among similar tokens have been introduced. However, existing approaches neglect the denoising priors of the diffusion models, leading to suboptimal acceleration and diminished image quality. This study proposes a novel concept: attend to prune feature redundancies in areas not attended by the diffusion process. We analyze the location and degree of feature redundancies based on the structure-then-detail denoising priors. Subsequently, we introduce SDTM, a structure-then-detail token merging approach that dynamically compresses feature redundancies. Specifically, we design dynamic visual token merg

ing, compression ratio adjusting, and prompt reweighting for different stages. S erved in a post-training way, the proposed method can be integrated seamlessly i nto any DiT architecture. Extensive experiments across various backbones, schedu lers, and datasets showcase the superiority of our method, for example, it achie ves 1.55 times acceleration with negligible impact on image quality. Project pag e: https://github.com/ICTMCG/SDTM.
****

Turbo3D: Ultra-fast Text-to-3D Generation
Hanzhe Hu, Tianwei Yin, Fujun Luan, Yiwei Hu, Hao Tan, Zexiang Xu, Sai Bi, Shubh am Tulsiani, Kai Zhang; Proceedings of the Computer Vision and Pattern Recogniti on Conference (CVPR), 2025, pp. 23668-23678
We present Turbo3D, an ultra-fast text-to-3D system capable of generating high-q uality Gaussian splatting assets in under one second. Turbo3D employs a rapid 4- step, 4-view diffusion generator, and an efficient feed-forward Gaussian reconst ructor, both operating in latent space. The 4-step, 4-view generator is a studen t model distilled through a novel Dual-Teacher approach, which encourages the st udent to learn view consistency from a multi-view teacher and photo-realism from a single-view teacher. By shifting the Gaussian reconstructor's inputs from pix el space to latent space, we eliminate the extra image decoding time and halve t he transformer sequence length for maximum efficiency. Our method demonstrates s uperior 3D generation results compared to previous baselines, while operating in a fraction of their runtime.
****

SUM Parts: Benchmarking Part-Level Semantic Segmentation of Urban Meshes
Weixiao Gao, Liangliang Nan, Hugo Ledoux; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24474-24484
Semantic segmentation in urban scene analysis has mainly focused on images or po int clouds, while textured meshes--offering richer spatial representation--remai n underexplored. This paper introduces SUM Parts, the first large-scale dataset for urban textured meshes with part-level semantic labels, covering about 2.5km^ 2 with 21 classes. The dataset was created using our designed annotation tool, s upporting both face and texture-based annotations with efficient interactive sel ection. We also provide a comprehensive evaluation of 3D semantic segmentation a nd interactive annotation methods on this dataset.
****

One-for-More: Continual Diffusion Model for Anomaly Detection
Xiaofan Li, Xin Tan, Zhuo Chen, Zhizhong Zhang, Ruixin Zhang, Rizen Guo, Guanna Jiang, Yulong Chen, Yanyun Qu, Lizhuang Ma, Yuan Xie; Proceedings of the Compute r Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4766-4775
With the rise of generative models, there is a growing interest in unifying all tasks within a generative framework. Anomaly detection methods also fall into th is scope and utilize diffusion models to generate or reconstruct normal samples when given arbitrary anomaly images. However, our study found that the diffusion model suffers from severe faithfulness hallucination and catastrophic forgettin g, which can't meet the unpredictable pattern increments. To mitigate the above problems, we propose a continual diffusion model that uses gradient projection t o achieve stable continual learning. Gradient projection deploys a regularizatio n on the model updating by modifying the gradient towards the direction protecti ng the learned knowledge. But as a double-edged sword, it also requires huge mem ory costs brought by the Markov process. Hence, we propose an iterative singular value decomposition method based on the transitive property of linear represent ation, which consumes tiny memory and incurs almost no performance loss. Finally , considering the risk of over-fitting to normal images of the diffusion model, we propose an anomaly-masked network to enhance the condition mechanism of the d iffusion model. For continual anomaly detection, ours achieves first place in 17 /18 settings on MVTec and VisA.
****

MODA: Motion-Drift Augmentation for Inertial Human Motion Analysis
Yinghao Wu, Shihui Guo, Yipeng Qin; Proceedings of the Computer Vision and Patte rn Recognition Conference (CVPR), 2025, pp. 27771-27781

While data augmentation (DA) has been extensively studied in computer vision, it s application to Inertial Measurement Unit (IMU) signals remains largely unexplo red, despite IMUs' growing importance in human motion analysis. In this paper, w e present the first systematic study of IMU-specific data augmentation, beginnin g with a comprehensive analysis that identifies three fundamental properties of IMU signals: their time-series nature, inherent multimodality (rotation and acce leration) and motion-consistency characteristics. Through this analysis, we demo nstrate the limitations of applying conventional time-series augmentation techni ques to IMU data. We then introduce Motion-Drift Augmentation (MODA), a novel te chnique that simulates the natural displacement of body-worn IMUs during motion. We evaluate our approach across five diverse datasets and five deep learning se ttings, including i) fully-supervised, ii) semi-supervised, iii) domain adaptati on, iv) domain generalization and v) few-shot learning for both Human Action Rec ognition (HAR) and Human Pose Estimation (HPE) tasks. Experimental results show that our proposed MODA consistently outperforms existing augmentation methods, w ith semi-supervised learning performance approaching state-of-the-art fully-supe rvised methods.

****

Higher-Order Ratio Cycles for Fast and Globally Optimal Shape Matching

Paul Roetzer, Viktoria Ehm, Daniel Cremers, Zorah Lähner, Florian Bernard; Proce edings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, p p. 21793-21803

In this work we address various shape matching problems that can be cast as find ing cyclic paths in a product graph. This involves for example 2D-3D shape match ing, 3D shape matching, or the matching of a contour to a graph. In this context , matchings are typically obtained as the minimum cost cycle in the product grap h. Instead, inspired by related works on model-based image segmentation (Schoene mann and Cremers 2009), we consider minimum ratio cycles, which we combine with the recently introduced conjugate product graph in order to allow for higher-ord er matching costs. With that, on the one hand we avoid the bias of obtaining mat chings that involve fewer/shorter edges, while on the other hand we are able to impose powerful geometric regularisation, e.g. to avoid zig-zagging. In our expe riments we demonstrate that this not only leads to improved matching accuracy in most cases, but also to significantly reduced runtimes (up to two orders of mag nitude, depending on the setting). Our GPU implementations are publicly availabl e: https://github.com/paul0noah/product-graph-cycles/.

****

Omni-RGPT: Unifying Image and Video Region-level Understanding via Token Marks

Miran Heo, Min-Hung Chen, De-An Huang, Sifei Liu, Subhashree Radhakrishnan, Seon Joo Kim, Yu-Chiang Frank Wang, Ryo Hachiuma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3919-3930

We present Omni-RGPT, a multimodal large language model designed to facilitate r egion-level comprehension for both images and videos. To achieve consistent regi on representation across spatio-temporal dimensions, we introduce Token Mark, a set of tokens highlighting the target regions within the visual feature space. T hese tokens are directly embedded into spatial regions using region prompts (e.g ., boxes or masks) and simultaneously incorporated into the text prompt to speci fy the target, establishing a direct connection between visual and text tokens. To further support robust video understanding without requiring tracklets, we in troduce an auxiliary task that guides Token Mark by leveraging the consistency o f the tokens, enabling stable region interpretation across the video. Additional ly, we introduce a large-scale region-level video instruction dataset (RegVID-30 0k). Omni-RGPT achieves state-of-the-art results on image and video-based common sense reasoning benchmarks while showing strong performance in captioning and re ferring expression comprehension tasks.

****

Hyperdimensional Uncertainty Quantification for Multimodal Uncertainty Fusion in Autonomous Vehicles Perception

Luke Chen, Junyao Wang, Trier Mortlock, Pramod Khargonekar, Mohammad Abdullah Al Faruque; Proceedings of the Computer Vision and Pattern Recognition Conference

(CVPR), 2025, pp. 22306-22316
Uncertainty Quantification (UQ) is crucial for ensuring the reliability of machine learning models deployed in real-world autonomous systems. However, existing approaches typically quantify task-level output prediction uncertainty without considering epistemic uncertainty at the multimodal feature fusion level, leading to sub-optimal outcomes. Additionally, popular uncertainty quantification methods, e.g., Bayesian approximations, remain challenging to deploy in practice due to high computational costs in training and inference. In this paper, we propose HyperDUM, a novel deterministic uncertainty method (DUM) that efficiently quantifies feature-level epistemic uncertainty by leveraging hyperdimensional computing. Our method captures the channel and spatial uncertainties through channel and patch -wise projection and bundling techniques respectively. Multimodal sensor features are then adaptively weighted to mitigate uncertainty propagation and improve feature fusion. Our evaluations show that HyperDUM on average outperforms the state-of-the-art (SOTA) algorithms by up to 2.01%/1.27% in 3D Object Detection and up to 1.29% improvement over baselines in semantic segmentation tasks under various types of uncertainties. Notably, HyperDUM requires 2.36x less Floating Point Operations and up to 38.30x less parameters than SOTA methods, providing an efficient solution for real-world autonomous systems.
****
EDM: Equirectangular Projection-Oriented Dense Kernelized Feature Matching
Dongki Jung, Jaehoon Choi, Yonghan Lee, Somi Jeong, Taejae Lee, Dinesh Manocha, Suyong Yeon; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6337-6347
We introduce the first learning-based dense matching algorithm, termed Equirectangular Projection-Oriented Dense Kernelized Feature Matching (EDM), specifically designed for omnidirectional images. Equirectangular projection (ERP) images, with their large fields of view, are particularly suited for dense matching techniques that aim to establish comprehensive correspondences across images. However, ERP images are subject to significant distortions, which we address by leveraging the spherical camera model and geodesic flow refinement in the dense matching method. To further mitigate these distortions, we propose spherical positional embeddings based on 3D Cartesian coordinates of the feature grid. Additionally, our method incorporates bidirectional transformations between spherical and Cartesian coordinate systems during refinement, utilizing a unit sphere to improve matching performance. We demonstrate that our proposed method achieves notable performance enhancements, with improvements of +26.72 and +42.62 in AUC@5deg on the Matterport3D and Stanford2D3D datasets. Project page: https://jdk9405.github.io/EDM
****
EZSR: Event-based Zero-Shot Recognition
Yan Yang, Liyuan Pan, Dongxu Li, Liu Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4628-4638
This paper studies zero-shot object recognition using event camera data. Guided by CLIP, which is pre-trained on RGB images, existing approaches achieve zero-shot object recognition by optimizing embedding similarities between event data and RGB images respectively encoded by an event encoder and the CLIP image encoder. Alternatively, several methods learn RGB frame reconstructions from event data for the CLIP image encoder. However, they often result in suboptimal zero-shot performance. This study develops an event encoder without relying on additional reconstruction networks. We theoretically analyze the performance bottlenecks of previous approaches: the embedding optimization objectives are prone to suffer from the spatial sparsity of event data, causing semantic misalignments between the learned event embedding space and the CLIP text embedding space. To mitigate the issue, we explore a scalar-wise modulation strategy. Furthermore, to scale up the number of events and RGB data pairs for training, we also study a pipeline for synthesizing event data from static RGB images in mass. Experimentally, we demonstrate an attractive scaling property in the number of parameters and synthesized data. We achieve superior zero-shot object recognition performance on extensive standard benchmark datasets, even compared with past supervised learning

approaches. For example, our model with a ViT/B-16 backbone achieves 47.84% zero-shot accuracy on the N-ImageNet dataset.

****

FlowRAM: Grounding Flow Matching Policy with Region-Aware Mamba Framework for Robotic Manipulation

Sen Wang, Le Wang, Sanping Zhou, Jingyi Tian, Jiayi Li, Haowen Sun, Wei Tang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12176-12186

Robotic manipulation in high-precision tasks is essential for numerous industrial and real-world applications where accuracy and speed are required. Yet current diffusion-based policy learning methods generally suffer from low computational efficiency due to the iterative denoising process during inference. Moreover, these methods do not fully explore the potential of generative models for enhancing information exploration in 3D environments. In response, we propose FlowRAM, a novel framework that leverages generative models to achieve region-aware perception, enabling efficient multimodal information processing. Specifically, we devise a Dynamic Radius Schedule, which enables adaptive perception, facilitating transitions from global scene comprehension to fine-grained geometric details. Furthermore, we integrate state space models to integrate multimodal information, while preserving linear computational complexity. In addition, we employ conditional flow matching to learn action poses by regressing deterministic vector fields, which simplifies the learning process while maintaining performance. We verify the effectiveness of the FlowRAM in the RLBench, an established manipulation benchmark, and achieve state-of-the-art performance. The results demonstrate that FlowRAM achieves a remarkable improvement, particularly in high-precision tasks, where it outperforms previous methods by 12.0% in average success rate. Additionally, FlowRAM is able to generate physically plausible actions for a variety of real-world tasks in less than 4 time steps, significantly increasing inference speed.

****

Visual Lexicon: Rich Image Features in Language Space

XuDong Wang, Xingyi Zhou, Alireza Fathi, Trevor Darrell, Cordelia Schmid; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19736-19747

We present Visual Lexicon, a novel visual language that encodes rich image information into the text space of vocabulary tokens while retaining intricate visual details that are often challenging to convey in natural language. Unlike traditional methods that prioritize either high-level semantics (e.g., CLIP) or pixel-level reconstruction (e.g., VAE), ViLex simultaneously captures rich semantic content and fine visual details, enabling high-quality image generation and comprehensive visual scene understanding. Through a self-supervised learning pipeline, ViLex generates tokens optimized for reconstructing input images using a frozen text-to-image (T2I) diffusion model, preserving the detailed information necessary for high-fidelity semantic-level reconstruction. As an image embedding in the language space, ViLex tokens leverage the compositionality of natural languages, allowing them to be used independently as "text tokens" or combined with natural language tokens to prompt pretrained T2I models with both visual and textual inputs, mirroring how we interact with vision-language models (VLMs). Experiments demonstrate that ViLex achieves higher fidelity in image reconstruction compared to text embeddings--even with a single ViLex token. Moreover, ViLex successfully performs various DreamBooth tasks in a zero-shot, unsupervised manner without fine-tuning T2I models. Additionally, ViLex serves as a powerful vision encoder, consistently improving vision-language model performance across 15 benchmarks relative to a strong SigLIP baseline.

****

SVFR: A Unified Framework for Generalized Video Face Restoration

Zhiyao Wang, Xu Chen, Chengming Xu, Junwei Zhu, Xiaobin Hu, Jiangning Zhang, Chengjie Wang, Yuqi Liu, Yiyi Zhou, Rongrong Ji; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7406-7415

Face Restoration (FR) is a crucial area within image and video processing, focus

ing on reconstructing high-quality portraits from degraded inputs. Despite advan cements in image FR, video FR remains relatively under-explored, primarily due t o challenges related to temporal consistency, motion artifacts, and the limited availability of high-quality video data. Moreover, traditional face restoration typically prioritizes enhancing resolution and may not give as much consideratio n to related tasks such as facial colorization and inpainting. In this paper, we propose a novel approach for the Generalized Video Face Restoration (GVFR) task , which integrates video blind face restoration (BFR), inpainting, and colorizat ion tasks that we empirically show to benefit each other. We present a unified f ramework, termed as stable video face restoration (SVFR), which leverages the ge nerative and motion priors of Stable Video Diffusion (SVD) and incorporates task -specific information through a unified face restoration framework. A learnable task embedding is introduced to enhance task identification. Meanwhile, a novel Unified Latent Regularization (ULR) is employed to encourage the shared feature representation learning among different subtasks. To further enhance the restora tion quality and temporal stability, we introduce the facial prior learning and the self-referred refinement as auxiliary strategies. The proposed framework eff ectively combines the complementary strengths of these tasks, enhancing temporal coherence and achieving superior restoration quality. This work advances the st ate-of-the-art in video FR and establishes a new paradigm for generalized video face restoration. Code and video demo are available at https://github.com/wangzh iyaoo/SVFR.git.

****

Decoupling Fine Detail and Global Geometry for Compressed Depth Map Super-Resolu tion

Huan Zheng, Wencheng Han, Jianbing Shen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 951-960

Recovering high-quality depth maps from compressed sources has gained significan t attention due to the limitations of consumer-grade depth cameras and the bandw idth restrictions during data transmission. However, current methods still suffe r from two challenges. First, bit-depth compression produces a uniform depth rep resentation in regions with subtle variations, hindering the recovery of detaile d information. Second, densely distributed random noise reduces the accuracy of estimating the global geometric structure of the scene. To address these challen ges, we propose a novel framework, termed geometry-decoupled network (GDNet), fo r compressed depth map super-resolution that decouples the high-quality depth ma p reconstruction process by handling global and detailed geometric features sepa rately. To be specific, we propose the fine geometry detail encoder (FGDE), whic h is designed to aggregate fine geometry details in high-resolution low-level im age features while simultaneously enriching them with complementary information from low-resolution context-level image features. In addition, we develop the gl obal geometry encoder (GGE) that aims at suppressing noise and extracting global geometric information effectively via constructing compact feature representati on in a low-rank space. We conduct experiments on multiple benchmark datasets, d emonstrating that our GDNet significantly outperforms current methods in terms o f geometric consistency and detail recovery. In the ECCV 2024 AIM Compressed Dep th Upsampling Challenge, our solution won the 1st place award. Our codes are ava ilable at: https://github.com/Ian0926/GDNet.

****

Test-Time Visual In-Context Tuning

Jiahao Xie, Alessio Tonioni, Nathalie Rauschmayr, Federico Tombari, Bernt Schiel e; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19996-20005

Visual in-context learning (VICL), as a new paradigm in computer vision, allows the model to rapidly adapt to various tasks with only a handful of prompts and e xamples. While effective, the existing VICL paradigm exhibits poor generalizabil ity under distribution shifts. In this work, we propose test-time Visual In-Cont ext Tuning (VICT), a method that can adapt VICL models on the fly with a single test sample. Specifically, we flip the role between the task prompts and the tes t sample and use a cycle consistency loss to reconstruct the original task promp

t output. Our key insight is that a model should be aware of a new test distribu tion if it can successfully recover the original task prompts. Extensive experim ents on six representative vision tasks ranging from high-level visual understan ding to low-level image processing, with 15 common corruptions, demonstrate that our VICT can improve the generalizability of VICL to unseen new domains. In add ition, we show the potential of applying VICT for unseen tasks at test time. Cod e: https://github.com/Jiahao000/VICT.
****

Prior Does Matter: Visual Navigation via Denoising Diffusion Bridge Models
Hao Ren, Yiming Zeng, Zetong Bi, Zhaoliang Wan, Junlong Huang, Hui Cheng; Procee dings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp . 12100-12110
Recent advancements in diffusion-based imitation learning, which shows impressiv e performance in modeling multimodal distributions and training stability, have led to substantial progress in various robot learning tasks. In visual navigatio n, previous diffusion-based policies typically generate action sequences by init iating from denoising Gaussian noise. However, the target action distribution of ten diverges significantly from Gaussian noise, leading to redundant denoising s teps and increased learning complexity. Additionally, the sparsity of effective action distributions makes it challenging for the policy to generate accurate ac tions without guidance. To address these issues, we propose a novel, unified vis ual navigation framework leveraging the denoising diffusion bridge models named NaviBridger. This approach enables action generation by initiating from any info rmative prior actions, enhancing guidance and efficiency in the denoising proces s. We explore how diffusion bridges can enhance imitation learning in visual nav igation tasks and further examine three source policies for generating prior act ions. Extensive experiments in both simulated and real-world indoor and outdoor scenarios demonstrate that NaviBridger accelerates policy inference and outperfo rms the baselines in generating target action sequences. Code is available at ht tps://github.com/hren20/NaiviBridger.
****

Digital Twin Catalog: A Large-Scale Photorealistic 3D Object Digital Twin Datase t
Zhao Dong, Ka Chen, Zhaoyang Lv, Hong-Xing Yu, Yunzhi Zhang, Cheng Zhang, Yufeng Zhu, Stephen Tian, Zhengqin Li, Geordie Moffatt, Sean Christofferson, James For t, Xiaqing Pan, Mingfei Yan, Jiajun Wu, Carl Yuheng Ren, Richard Newcombe; Proce edings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, p p. 753-763
We introduce Digital Twin Catalog (DTC), a new large-scale photorealistic 3D obj ect digital twin dataset. A digital twin of a 3D object is a highly detailed, vi rtually indistinguishable representation of a physical object, accurately captur ing its shape, appearance, physical properties, and other attributes. Recent adv ances in neural-based 3D reconstruction and inverse rendering have significantly improved the quality of 3D object reconstruction. Despite these advancements, t here remains a lack of a large-scale, digital twin quality real-world dataset an d benchmark that can quantitatively assess and compare the performance of differ ent reconstruction methods, as well as improve reconstruction quality through tr aining or fine-tuning. Moreover, to democratize 3D digital twin creation, it is essential to integrate creation techniques with next-generation egocentric compu ting platforms, such as AR glasses. Currently, there is no dataset available to evaluate 3D object reconstruction using egocentric captured images. To address t hese gaps, the DTC dataset features 2,000 scanned digital twin-quality 3D object s, along with image sequences captured under different lighting conditions using DSLR cameras and egocentric AR glasses. This dataset establishes the first comp rehensive real-world evaluation benchmark for 3D digital twin creation tasks, of fering a robust foundation for comparing and improving existing reconstruction m ethods. The DTC dataset is already released at https://www.projectaria.com/datas ets/dtc/ and we will also make the baseline evaluations open-source.
****

SegEarth-OV: Towards Training-Free Open-Vocabulary Segmentation for Remote Sensi

ng Images
Kaiyu Li, Ruixun Liu, Xiangyong Cao, Xueru Bai, Feng Zhou, Deyu Meng, Zhi Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10545-10556
Current remote sensing semantic segmentation methods are mostly built on the close-set assumption, meaning that the model can only recognize pre-defined categories that exist in the training set. However, in practical Earth observation, there are countless unseen categories, and manual annotation is impractical. To address this challenge, we first attempt to introduce training-free open-vocabulary semantic segmentation (OVSS) into the remote sensing context. However, due to the sensitivity of remote sensing images to low-resolution features, distorted target shapes and ill-fitting boundaries are exhibited in the prediction mask. To tackle these issues, we propose a simple and universal upsampler, i.e. SimFeatUp, to restore lost spatial information of deep features. Specifically, SimFeatUp only needs to learn from a few unlabeled images, and can upsample arbitrary remote sensing image features. Furthermore, based on the observation of the abnormal response of patch tokens to the [CLS] token in CLIP, we propose to execute a simple subtraction operation to alleviate the global bias in patch tokens. Extensive experiments are conducted on 17 remote sensing datasets of 4 tasks, including semantic segmentation, building extraction, road detection, and flood detection. Our method achieves an average of 5.8%, 8.2%, 4.0%, and 15.3% improvement over state-of-the-art methods on the 4 tasks.
****

MeshGen: Generating PBR Textured Mesh with Render-Enhanced Auto-Encoder and Generative Data Augmentation
Zilong Chen, Yikai Wang, Wenqiang Sun, Feng Wang, Yiwen Chen, Huaping Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5835-5848
In this paper, we introduce MeshGen, an advanced image-to-3D pipeline that generates high-quality 3D meshes with detailed geometry and physically based rendering (PBR) textures. Addressing the challenges faced by existing 3D native diffusion models, such as suboptimal auto-encoder performance, limited controllability, poor generalization, and inconsistent image-based PBR texturing, MeshGen employs several key innovations to overcome these limitations. We pioneer a render-enhanced point-to-shape auto-encoder that compresses meshes into a compact latent space, by designing perceptual optimization with ray-based regularization. This ensures that the 3D shapes are accurately represented and reconstructed to preserve geometric details within the latent space. To address data scarcity and image-shape misalignment, we further propose geometric augmentation and generative rendering augmentation techniques, which enhance the model's controllability and generalization ability, allowing it to perform well even with limited public datasets. For texture generation, MeshGen employs a reference attention-based multi-view ControlNet for consistent appearance synthesis. This is further complemented by our multi-view PBR decomposer that estimates PBR components and a UV inpainter that fills invisible areas, ensuring a seamless and consistent texture across the 3D mesh. Our extensive experiments demonstrate that MeshGen largely outperforms previous methods in both shape and texture generation, setting a new standard for the quality of 3D meshes generated with PBR textures.
****

GIFStream: 4D Gaussian-based Immersive Video with Feature Stream
Hao Li, Sicheng Li, Xiang Gao, Abudouaihati Batuer, Lu Yu, Yiyi Liao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21761-21770
Immersive video offers a 6-Dof-free viewing experience, potentially playing a key role in future video technology. Recently, 4D Gaussian Splatting has gained attention as an effective approach for immersive video due to its high rendering efficiency and quality, though maintaining quality with manageable storage remains challenging. To address this, we introduce GIFStream, a novel 4D Gaussian representation using a canonical space and a deformation field enhanced with time-dependent feature streams. These feature streams enable complex motion modeling an

d allow efficient compression by leveraging their motion-awareness and temporal correspondence. Additionally, we incorporate both temporal and spatial compression networks for end-to-end compression. Experimental results show that GIFStream delivers high-quality immersive video at 30 Mbps, with real-time rendering and fast decoding on an RTX 4090.

****

DeClotH: Decomposable 3D Cloth and Human Body Reconstruction from a Single Image
Hyeongjin Nam, Donghwan Kim, Jeongtaek Oh, Kyoung Mu Lee; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5636-5645
Most existing methods of 3D clothed human reconstruction from a single image treat the clothed human as a single object without distinguishing between cloth and human body. In this regard, we present DeClotH, which separately reconstructs 3D cloth and human body from a single image. This task remains largely unexplored due to the extreme occlusion between cloth and the human body, making it challenging to infer accurate geometries and textures. Moreover, while recent 3D human reconstruction methods have achieved impressive results using text-to-image diffusion models, directly applying such an approach to this problem often leads to incorrect guidance, particularly in reconstructing 3D cloth. To address these challenges, we propose two core designs in our framework. First, to alleviate the occlusion issue, we leverage 3D template models of cloth and human body as regularizations, which provide strong geometric priors to prevent erroneous reconstruction by the occlusion. Second, we introduce a cloth diffusion model specifically designed to provide contextual information about cloth appearance, thereby enhancing the reconstruction of 3D cloth. Qualitative and quantitative experiments demonstrate that our proposed approach is highly effective in reconstructing both 3D cloth and the human body.

****

Neuron: Learning Context-Aware Evolving Representations for Zero-Shot Skeleton Action Recognition
Yang Chen, Jingcai Guo, Song Guo, Dacheng Tao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8721-8730
Zero-shot skeleton action recognition is a non-trivial task that requires robust unseen generalization with prior knowledge from only seen classes and shared semantics. Existing methods typically build the skeleton-semantics interactions by uncontrollable mappings and conspicuous representations, thereby can hardly capture the intricate and fine-grained relationship for effective cross-modal transferability. To address these issues, we propose a novel dyNamically Evolving dUal skeleton-semantic syneRgistic framework with the guidance of cOntext-aware side informatioN (dubbed Neuron), to explore more fine-grained cross-modal correspondence from micro to macro perspectives at both spatial and temporal levels, respectively. Concretely, 1) we first construct the spatial-temporal evolving micro-prototypes and integrate dynamic context-aware side information to capture the intricate and synergistic skeleton-semantic correlations step-by-step, progressively refining cross-model alignment; and 2) we introduce the spatial compression and temporal memory mechanisms to guide the growth of spatial-temporal micro-prototypes, enabling them to absorb structure-related spatial representations and regularity-dependent temporal patterns. Notably, such processes are analogous to the learning and growth of neurons, equipping the framework with the capacity to generalize to novel unseen action categories. Extensive experiments on various benchmark datasets demonstrated the superiority of the proposed method. Code is available at: https://github.com/cseeyangchen/Neuron.

****

Multi-Scale Neighborhood Occupancy Masked Autoencoder for Self-Supervised Learning in LiDAR Point Clouds
Mohamed Abdelsamad, Michael Ulrich, Claudius Glaeser, Abhinav Valada; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22234-22243
Masked autoencoders (MAE) have shown tremendous potential for self-supervised learning (SSL) in vision and beyond. However, point clouds from LiDARs used in automated driving are particularly challenging for MAEs since large areas of the 3D

volume are empty. Consequently, existing work suffers from leaking occupancy in
formation into the decoder and has significant computational complexity, thereby
limiting the SSL pre-training to only 2D bird's eye view encoders in practice.
In this work, we propose the novel neighborhood occupancy MAE (NOMAE) that overc
omes the aforementioned challenges by employing masked occupancy reconstruction
only in the neighborhood of non-masked voxels. We incorporate voxel masking and
occupancy reconstruction at multiple scales with our proposed hierarchical mask
generation technique to capture features of objects of different sizes in the po
int cloud. NOMAEs are extremely flexible and can be directly employed for SSL in
existing 3D architectures. We perform extensive evaluations on the nuScenes and
Waymo Open datasets for the downstream perception tasks of semantic segmentatio
n and 3D object detection, comparing with both discriminative and generative SSL
methods. The results demonstrate that NOMAE sets the new state-of-the-art on mu
ltiple benchmarks for multiple point cloud perception tasks.
****

Do We Really Need Curated Malicious Data for Safety Alignment in Multi-modal Lar
ge Language Models?
Yanbo Wang, Jiyang Guan, Jian Liang, Ran He; Proceedings of the Computer Vision
and Pattern Recognition Conference (CVPR), 2025, pp. 19879-19889
Multi-modal large language models (MLLMs) have made significant progress, yet th
eir safety alignment remains limited. Typically, current open-source MLLMs rely
on the alignment inherited from their language module to avoid harmful generatio
ns. However, the lack of safety measures specifically designed for multi-modal i
nputs creates an alignment gap, leaving MLLMs vulnerable to vision-domain attack
s such as typographic manipulation. Current methods utilize a carefully designed
safety dataset to enhance model defense capability.However, it is unknown what
is actually learned in the high-quality dataset.Through comparison experiments,
we find that the alignment gap primarily arises from data distribution biases, w
hile image content, response quality, or the contrastive behavior of the dataset
makes little contribution to boosting multi-modal safety. To further investigat
e this and identify the key factors in improving MLLM safety, we propose finetun
ing MLLMs on a small set of benign instruct-following data with responses replac
ed by simple, clear rejection sentences.Experiments show that, without the need
for labor-intensive collection of high-quality malicious data, model safety can
still be significantly improved, as long as a specific fraction of rejection dat
a exists in the finetuning set, indicating the security alignment is not lost bu
t rather obscured during multi-modal pretraining or instruction fine-tuning. Sim
ply correcting the underlying data bias is enough to address the vision domain s
afety gap.
****

High-Fidelity Relightable Monocular Portrait Animation with Lighting-Controllabl
e Video Diffusion Model
Mingtao Guo, Guanyu Xing, Yanli Liu; Proceedings of the Computer Vision and Patt
ern Recognition Conference (CVPR), 2025, pp. 228-238
Relightable portrait animation aims to animate a static reference portrait to ma
tch the head movements of a driving video while adapting to user-specified or re
ference lighting conditions. Existing portrait animation methods fail to achieve
relightable portraits because they do not separate and manipulate intrinsic (id
entity and appearance) and extrinsic (pose and lighting) features. In this paper
, we present a Lighting Controllable Video Diffusion model (LCVD) for high-fidel
ity, relightable portrait animation. We address this limitation by distinguishin
g these feature types through dedicated subspaces within the feature space of a
pre-trained image-to-video diffusion model. Specifically, for each frame, we use
the 3D mesh of the reference portrait, the pose of the driving frame as well as
the specified lighting to render an image called shading hint. While the refere
nce image represents the intrinsic attributes, the shading hint encodes the extr
insic attributes. In the training phase, we employ a reference adapter to map th
e reference into an intrinsic feature subspace and a shading adapter to map the
shading hints into an extrinsic feature subspace. By merging features from these
subspaces, the model achieves nuanced control over lighting and pose in generat

ed animations. Extensive evaluations show that LCVD outperforms state-of-the-art methods in lighting realism, image quality, and video consistency, setting a new benchmark in relightable portrait animation. Our project is available at \href https://github.com/MingtaoGuo/Relightable-Portrait-Animation https://github.com/MingtaoGuo/Relightable-Portrait-Animation
****

Plug-and-Play PPO: An Adaptive Point Prompt Optimizer Making SAM Greater
Xueyu Liu, Rui Wang, Yexin Lai, Guangze Shi, Feixue Shao, Fang Hao, Jianan Zhang, Jia Shen, Yongfei Wu, Wen Zheng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4332-4342
Powered by extensive curated training data, the Segment Anything Model (SAM) demonstrates impressive generalization capabilities in open-world scenarios, effectively guided by user-provided prompts. However, the class-agnostic characteristic of SAM renders its segmentation accuracy highly dependent on prompt quality. In this paper, we propose a novel pplug-and-play dual-space Point Prompt Optimizer (PPO) designed to enhance prompt distribution through deep reinforcement learning (DRL)-based heterogeneous graph optimization. PPO optimizes initial prompts for any task without requiring additional training, thereby improving SAM's downstream segmentation performance. Specifically, PPO constructs a dual-space heterogeneous graph, leveraging the robust feature-matching capabilities of a foundational pre-trained model to create internal feature and physical distance matrices. A DRL policy network iteratively refines the distribution of prompt points, optimizing segmentation predictions. We conducted experiments on four public datasets. The ablation study explores the necessity and balance of optimizing prompts in both feature and physical spaces. The comparative study shows that PPO enables SAM to surpass recent one-shot methods. Additionally, experiments with different initial prompts demonstrate PPO's generality across prompts generated by various methods. In conclusion, PPO redefines the prompt optimization problem as a heterogeneous graph optimization task, using DRL to construct an effective, plug-and-play prompt optimizer. This approach holds potential for broader applications across diverse segmentation tasks and provides a promising solution for point prompt optimization. The source code and demo are available at https://github.com/XueyuLiu/PPO.
****

Harnessing Global-Local Collaborative Adversarial Perturbation for Anti-Customization
Long Xu, Jiakai Wang, Haojie Hao, Haotong Qin, Jiejie Zhao, Xianglong Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13414-13423
Though achieving significant success in personalized image synthesis, Latent Diffusion Models (LDMs) pose substantial social risks caused by unauthorized misuse (e.g., face theft). To counter these threats, the Anti-Customization (AC) method that exploits adversarial perturbations was proposed. Unfortunately, existing AC methods show insufficient defense ability due to the ignorance of hierarchical characteristics, i.e., global feature correlations and local facial attributes, leading to weak resistance to concept transfer and semantic theft in customization methods. To address these limitations, we are motivated to propose a **G**lobal-L**o**cal c**o**llaborate**d** Anti-Customization (GoodAC) framework to generate powerful adversarial perturbations by disturbing both feature correlations and facial attributes. To enhance the ability to resist concept transfer, we disrupt the spatial correlation of perceptual features that form the basis of model generation at a global level, thereby creating highly concept-transfer-resistant adversarial camouflage. To improve the ability to resist semantic theft, leveraging the fact that facial attributes are personalized, we designed a personalized and precise facial attribute distortion strategy locally, focusing the attack on the individual's image structure to generate strong camouflage. Extensive experiments on various customization methods, including Dreambooth and LoRA, have strongly demonstrated that our GoodAC outperforms other state-of-the-art approaches by large margins, e.g., over 50% improvements on ISM.
****

EchoONE: Segmenting Multiple Echocardiography Planes in One Model
Jiongtong Hu, Wufeng Xue, Jun Cheng, Yingying Liu, Wei Zhuo, Dong Ni; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5207-5216
In clinical practice of echocardiography examinations, multiple planes containing the heart structures of different view are usually required in screening, diagnosis and treatment of cardiac disease. AI models for echocardiography have to be tailored for a specific plane due to the dramatic structure differences, thus resulting in repetition development and extra complexity. Effective solution for such a multi-plane segmentation (MPS) problem is highly demanded for medical images, yet has not been well investigated. In this paper, we propose a novel solution, EchoONE, for this problem with a SAM-based segmentation architecture, a prior-composable mask learning (PC-Mask) module for semantic-aware dense prompt generation, and a learnable CNN-branch with a simple yet effective local feature fusion and adaption (LFFA) module for SAM adapting. We extensively evaluated our method on multiple internal and external echocardiography datasets, and achieved consistently state-of-the-art performance for multi-source datasets with different heart planes. This is the first time that the MPS problem is solved in one model for echocardiography data. The code will be available at https://github.com/a2502503/EchoONE.
****
Acc3D: Accelerating Single Image to 3D Diffusion Models via Edge Consistency Guided Score Distillation
Kendong Liu, Zhiyu Zhu, Hui Liu, Junhui Hou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18031-18040
We present Acc3D to tackle the challenge of accelerating the diffusion process to generate 3D models from single images. To derive high-quality reconstructions through few-step inferences, we emphasize the critical issue of regularizing the learning of score function in states of random noise. To this end, we propose edge consistency, i.e., consistent predictions across the high signal-to-noise ratio region, to enhance a pre-trained diffusion model, enabling a distillation-based refinement of the endpoint score function. Building on those distilled diffusion models, we propose an adversarial augmentation strategy to further enrich the generation detail and boost overall generation quality. The two modules complement each other, mutually reinforcing to elevate generative performance. Extensive experiments demonstrate that our Acc3D not only achieves over a 20xincrease in computational efficiency but also yields notable quality improvements, compared to the state-of-the-arts.
****
EasyHOI: Unleashing the Power of Large Models for Reconstructing Hand-Object Interactions in the Wild
Yumeng Liu, Xiaoxiao Long, Zemin Yang, Yuan Liu, Marc Habermann, Christian Theobalt, Yuexin Ma, Wenping Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7037-7047
Our work aims to reconstruct hand-object interactions from a single-view image, which is a fundamental but ill-posed task.Unlike methods that reconstruct from videos, multi-view images, or predefined 3D templates, single-view reconstruction faces significant challenges due to inherent ambiguities and occlusions. These challenges are further amplified by the diverse nature of hand poses and the vast variety of object shapes and sizes. Our key insight is that current foundational models for segmentation, inpainting, and 3D reconstruction robustly generalize to in-the-wild images, which could provide strong visual and geometric priors for reconstructing hand-object interactions.Specifically, given a single image, we first design a novel pipeline to estimate the underlying hand pose and object shape using off-the-shelf large models. Furthermore, with the initial reconstruction, we employ a prior-guided optimization scheme, which optimizes hand pose to comply with 3D physical constraints and the 2D input image content.We perform experiments across several datasets and show that our method consistently outperforms baselines and faithfully reconstructs a diverse set of hand-object interactions. Here is the link of our project page: https://lym29.github.io/EasyHOI-pag

e/.
****
PLeaS - Merging Models with Permutations and Least Squares
Anshul Nasery, Jonathan Hayase, Pang Wei Koh, Sewoong Oh; Proceedings of the Com puter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30493-30502
The democratization of machine learning systems has made the process of fine-tun ing accessible to practitioners, leading to a wide range of open-source models f ine-tuned on specialized tasks and datasets. Recent work has proposed to merge s uch models to combine their functionalities. However, prior approaches are usual ly restricted to models that are fine-tuned from the same base model. Furthermor e, the final merged model is typically required to be of the same size as the or iginal models. In this work, we propose a new two-step algorithm to merge models ---termed PLeaS---which relaxes these constraints.First, leveraging the Permutat ion symmetries inherent in the two models, PLeaS partially matches nodes in each layer by maximizing alignment. Next, PLeaS computes the weights of the merged m odel as a layer-wise Least Squares solution to minimize the approximation error between the features of the merged model and the permuted features of the origin al models. PLeaS allows a practitioner to merge two models sharing the same arch itecture into a single performant model of a desired size, even when the two ori ginal models are fine-tuned from different base models. We also demonstrate how our method can be extended to address a challenging scenario where no data is av ailable from the fine-tuning domains. We demonstrate our method to merge ResNet and ViT models trained with shared and different label spaces, and show improvem ent over the state-of-the-art merging methods of up to 15 percentage points for the same target compute while merging models trained on DomainNet and fine-grai ned classification tasks.
****
Incremental Object Keypoint Learning
Mingfu Liang, Jiahuan Zhou, Xu Zou, Ying Wu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25399-25410
Existing progress in object keypoint estimation primarily benefits from the conv entional supervised learning paradigm based on numerous data labeled with pre-de fined keypoints. However, these well-trained models can hardly detect the undefi ned new keypoints in test time, which largely hinders their feasibility for dive rse downstream tasks. To handle this, various solutions are explored but still s uffer from either limited generalizability or transferability. Therefore, in thi s paper, we explore a novel keypoint learning paradigm in that we only annotate new keypoints in the new data and incrementally train the model, without retaini ng any old data, called Incremental object Keypoint Learning (IKL). A two-stage learning scheme as a novel baseline tailored to IKL is developed. In the first K nowledge Association stage, given the data labeled with only new keypoints, an a uxiliary KA-Net is trained to automatically associate the old keypoints to these new ones based on their spatial and intrinsic anatomical relations. In the seco nd Mutual Promotion stage, based on a keypoint-oriented spatial distillation los s, we jointly leverage the auxiliary KA-Net and the old model for knowledge cons olidation to mutually promote the estimation of all old and new keypoints. Owing to the investigation of the correlations between new and old keypoints, our pro posed method can not just effectively mitigate the catastrophic forgetting of ol d keypoints, but may even further improve the estimation of the old ones and ach ieve a positive transfer beyond anti-forgetting. Such an observation has been so lidly verified by extensive experiments on different keypoint datasets, where ou r method exhibits superiority in alleviating the forgetting issue and boosting p erformance while enjoying labeling efficiency even under the low-shot data regim e.
****
Soft Self-labeling and Potts Relaxations for Weakly-supervised Segmentation
Zhongwen Zhang, Yuri Boykov; Proceedings of the Computer Vision and Pattern Reco gnition Conference (CVPR), 2025, pp. 20244-20253
We consider weakly supervised segmentation where only a fraction of pixels have ground truth labels (scribbles) and focus on a self-labeling approach optimizing

relaxations of the standard unsupervised CRF/Potts loss on unlabeled pixels. While WSSS methods can directly optimize such losses via gradient descent, prior work suggests that higher-order optimization can improve network training by introducing hidden pseudo-labels and powerful CRF sub-problem solvers, e.g. graph cut. However, previously used hard pseudo-labels can not represent class uncertainty or errors, which motivates soft self-labeling. We derive a principled auxiliary loss and systematically evaluate standard and new CRF relaxations (convex and non-convex), neighborhood systems, and terms connecting network predictions with soft pseudo-labels. We also propose a general continuous sub-problem solver. Using only standard architectures, soft self-labeling consistently improves scribble-based training and outperforms significantly more complex specialized WSSS systems. It can outperform full pixel-precise supervision. Our general ideas apply to other weakly-supervised problems/systems.

****

MVSAnywhere: Zero-Shot Multi-View Stereo

Sergio Izquierdo, Mohamed Sayed, Michael Firman, Guillermo Garcia-Hernando, Daniyar Turmukhambetov, Javier Civera, Oisin Mac Aodha, Gabriel Brostow, Jamie Watson; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11493-11504

Computing accurate depth from multiple views is a fundamental and longstanding challenge in computer vision.However, most existing approaches do not generalize well across different domains and scene types (e.g. indoor vs outdoor). Training a general-purpose multi-view stereo model is challenging and raises several questions, e.g. how to best make use of transformer-based architectures, how to incorporate additional metadata when there is a variable number of input views, and how to estimate the range of valid depths which can vary considerably across different scenes and is typically not known a priori? To address these issues, we introduce MVSA, a novel and versatile Multi-View Stereo architecture that aims to work Anywhere by generalizing across diverse domains and depth ranges. MVSA combines monocular and multi-view cues with an adaptive cost volume to deal with scale-related issues. We demonstrate state-of-the-art zero-shot depth estimation on the Robust Multi-View Depth Benchmark, surpassing existing multi-view stereo and monocular baselines.

****

InteractVLM: 3D Interaction Reasoning from 2D Foundational Models

Sai Kumar Dwivedi, Dimitrije Anti■, Shashank Tripathi, Omid Taheri, Cordelia Schmid, Michael J. Black, Dimitrios Tzionas; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22605-22615

We introduce InteractVLM, a novel method to estimate 3D contact points on human bodies and objects from single in-the-wild images, enabling accurate human-object joint reconstruction in 3D. This is challenging due to occlusions, depth ambiguities, and widely varying object shapes. Existing methods rely on 3D contact annotations collected via expensive motion-capture systems or tedious manual labeling, limiting scalability and generalization. To overcome this, InteractVLM harnesses the broad visual knowledge of large Vision-Language Models (VLMs), fine-tuned with limited 3D contact data. However, directly applying these models is non-trivial, as they reason only in 2D, while human-object contact is inherently 3D. Thus we introduce a novel Render-Localize-Lift module that: (1) embeds 3D body and object surfaces in 2D space via multi-view rendering, (2) trains a novel multi-view localization model (MV-Loc) to infer contacts in 2D, and (3) lifts these to 3D. Additionally, we propose a new task called Semantic Human Contact estimation, where human contact predictions are conditioned explicitly on object semantics, enabling richer interaction modeling. InteractVLM outperforms existing work on contact estimation and also facilitates 3D reconstruction from an in-the wild image.

****

Patch Matters: Training-free Fine-grained Image Caption Enhancement via Local Perception

Ruotian Peng, Haiying He, Yake Wei, Yandong Wen, Di Hu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3963-3973

High-quality image captions play a crucial role in improving the performance of cross-modal applications such as text-to-image generation, text-to-video generation, and text-image retrieval. To generate long-form, high-quality captions, many recent studies have employed multimodal large language models (MLLMs). However, current MLLMs often produce captions that lack fine-grained details or suffer from hallucinations, a challenge that persists in both open-source and closed-source models. Inspired by Feature-Integration theory, which suggests that attention must focus on specific regions to integrate visual information effectively, we propose a divide-then-aggregate strategy. Our method first divides the image into semantic and spatial patches to extract fine-grained details, enhancing the model's local perception of the image. These local details are then hierarchically aggregated to generate a comprehensive global description. To address hallucinations and inconsistencies in the generated captions, we apply a semantic-level filtering process during hierarchical aggregation. This training-free pipeline can be applied to both open-source models (LLaVA-1.5, LLaVA-1.6, Mini-Gemini) and closed-source models (Claude-3.5-Sonnet, GPT-4o, GLM-4V-Plus). Extensive experiments demonstrate that our method generates more detailed, reliable captions, advancing multimodal description generation without requiring model retraining.

****

Attribute-Missing Multi-view Graph Clustering

Bowen Zhao, Qianqian Wang, Zhengming Ding, Quanxue Gao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25832-25841

The success of existing deep multi-view graph clustering methods is based on the assumption that node attributes are fully available across all views. However, in practical scenarios, node attributes are frequently missing due to factors such as data privacy concerns or failures in data collection devices. Although some methods have been proposed to address the issue of missing node attributes, they come with the following limitations: i) Existing methods are often not tailored specifically for clustering tasks and struggle to address missing attributes effectively. ii) They tend to ignore the relational dependencies between nodes and their neighboring nodes. This oversight results in unreliable imputations, thereby degrading clustering performance. To address the above issues, we propose an Attribute-Missing Multi-view Graph Clustering (AMMGC). Specifically, we first impute missing node attributes by leveraging neighborhood information through an adjacency matrix. Then, to improve the consistency, we integrate a dual structure consistency module that aligns graph structures across multiple views, reducing redundancy and retaining key information. Furthermore, we introduce a high-confidence guidance module to improve the reliability of clustering. Extensive experiment results showcase the effectiveness and superiority of our proposed method on multiple benchmark datasets.

****

Generating 6DoF Object Manipulation Trajectories from Action Description in Egocentric Vision

Tomoya Yoshida, Shuhei Kurita, Taichi Nishimura, Shinsuke Mori; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17370-17382

Learning to use tools or objects in common scenes, particularly handling them in various ways as instructed, is a key challenge for developing interactive robots. Training models to generate such manipulation trajectories requires a large and diverse collection of detailed manipulation demonstrations for various objects, which is nearly unfeasible to gather at scale. In this paper, we propose a framework that leverages large-scale ego- and exo-centric video datasets --- constructed globally with substantial effort --- of Exo-Ego4D to extract diverse manipulation trajectories at scale. From these extracted trajectories with the associated textual action description, we develop trajectory generation models based on visual and point cloud-based language models. In the recently proposed egocentric vision-based in-a-quality trajectory dataset of HOT3D, we confirmed that our models successfully generate valid object trajectories, establishing a training dataset and baseline models for the novel task of generating 6DoF manipulation trajectories from action descriptions in egocentric vision.

****
Pose-Guided Temporal Enhancement for Robust Low-Resolution Hand Reconstruction
Kaixin Fan, Pengfei Ren, Jingyu Wang, Haifeng Sun, Qi Qi, Zirui Zhuang, Jianxin Liao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22627-22637
3D hand reconstruction is essential in non-contact human-computer interaction applications, but existing methods struggle with low-resolution images, which occur in slightly distant interactive scenes. Leveraging temporal information can mitigate the limitations of individual low-resolution images that lack detailed appearance information, thereby enhancing the robustness and accuracy of hand reconstruction. Existing temporal methods typically use joint features to represent temporal information, avoiding interference from redundant background information. However, joint features excessively disregard the spatial context of visual features, limiting hand reconstruction accuracy. We propose to integrate temporal joint features with visual features to construct a robust low-resolution visual representation. We introduce Triplane Features, a dense representation with 3D spatial awareness, to bridge the gap between the joint features and visual features that are misaligned in terms of representation form and semantics. Triplane Features are obtained by orthogonally projecting the joint features, embedding hand structure information into the 3D spatial context. Furthermore, we compress the spatial information of the three planes into a 2D dense feature thourgh Spatial-Aware Fusion to enhance the visual features. By using enhanced visual features enriched with temporal information for hand reconstruction, our method achieves competitive performance at much lower resolutions compared to state-of-the-art methods operating at high resolution on DexYCB, HanCo and H2O.
****
ReDiffDet: Rotation-equivariant Diffusion Model for Oriented Object Detection
Jiaqi Zhao, Zeyu Ding, Yong Zhou, Hancheng Zhu, Wen-Liang Du, Rui Yao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24429-24439
The diffusion model has been successfully applied to various detection tasks. However, it still faces several challenges when used for oriented object detection: objects that are arbitrarily rotated require the diffusion model to encode their orientation information; uncontrollable random boxes inaccurately locate objects with dense arrangements and extreme aspect ratios; oriented boxes result in the misalignment between them and image features. To overcome these limitations, we propose ReDiffDet, a framework that formulates oriented object detection as a rotation-equivariant denoising diffusion process. First, we represent an oriented box as a 2D Gaussian distribution, forming the basis of the denoising paradigm. The reverse process can be proven to be rotation-equivariant within this representation and model framework. Second, we design a conditional encoder with conditional boxes to prevent boxes from being randomly placed across the entire image. Third, we propose an aligned decoder for alignment between oriented boxes and image features. The extensive experiments demonstrate ReDiffDet achieves promising performance and significantly outperforms the diffusion-based baseline detector.
****
PosterO: Structuring Layout Trees to Enable Language Models in Generalized Content-Aware Layout Generation
HsiaoYuan Hsu, Yuxin Peng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8117-8127
In poster design, content-aware layout generation is crucial for automatically arranging visual-textual elements on the given image. With limited training data, existing work focused on image-centric enhancement. However, this neglects the diversity of layouts and fails to cope with shape-variant elements or diverse design intents in generalized settings. To this end, we proposed a layout-centric approach that leverages layout knowledge implicit in large language models (LLMs) to create posters for omnifarious purposes, hence the name PosterO. Specifically, it structures layouts from datasets as trees in SVG language by universal shape, design intent vectorization, and hierarchical node representation. Then, it

applies LLMs during inference to predict new layout trees by in-context learning with intent-aligned example selection. After layout trees are generated, we can seamlessly realize them into poster designs by editing the chat with LLMs. Extensive experimental results have demonstrated that PosterO can generate visually appealing layouts for given images, achieving new state-of-the-art performance across various benchmarks. To further explore PosterO's abilities under the generalized settings, we built PStylish7, the first dataset with multi-purpose posters and various-shaped elements, further offering a challenging test for advanced research. Code and dataset are publicly available at Project page.
****

BIOMEDICA: An Open Biomedical Image-Caption Archive, Dataset, and Vision-Language Models Derived from Scientific Literature
Alejandro Lozano, Min Woo Sun, James Burgess, Liangyu Chen, Jeffrey J. Nirschl, Jeffrey Gu, Ivan Lopez, Josiah Aklilu, Anita Rau, Austin Wolfgang Katzer, Yuhui Zhang, Collin Chiu, Xiaohan Wang, Alfred Seunghoon Song, Robert Tibshirani, Serena Yeung-Levy; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19724-19735
The development of vision-language models (VLMs) is driven by large-scale and diverse multi-modal datasets. However, progress toward generalist biomedical VLMs is limited by the lack of annotated, publicly accessible datasets across biology and medicine. Existing efforts are limited to narrow domains, missing the opportunity to leverage the full diversity of biomedical knowledge encoded in scientific literature. To address this gap, we introduce BIOMEDICA: a scalable, open-source framework to extract, annotate, and serialize the entirety of the PubMed Central Open Access subset into an easy-to-use, publicly accessible dataset. Our framework produces a comprehensive archive with over 24 million unique image-text pairs from over 6 million articles. Metadata and expert-guided annotations are additionally provided. We demonstrate the utility and accessibility of our resource by releasing BMC-CLIP, a suite of CLIP-style models continuously pre-trained on BIOMEDICA dataset via streaming (eliminating the need to download 27 TB of data locally). On average, our models achieve state-of-the-art performance across 40 tasks -- spanning pathology, radiology, ophthalmology, dermatology, surgery, molecular biology, parasitology, and cell biology -- excelling in zero-shot classification with 6.56% average improvement (as high as 29.8% and 17.5% gains in dermatology and ophthalmology, respectively) and stronger image-text retrieval while using 10x less compute. To foster reproducibility and collaboration, we release our codebase and dataset to the broader research community
****

Unlocking Generalization Power in LiDAR Point Cloud Registration
Zhenxuan Zeng, Qiao Wu, Xiyu Zhang, Lin Yuanbo Wu, Pei An, Jiaqi Yang, Ji Wang, Peng Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22244-22253
In real-world environments, a LiDAR point cloud registration method with robust generalization capabilities (across varying distances and datasets) is crucial for ensuring safety in autonomous driving and other LiDAR-based applications. However, current methods fall short in achieving this level of generalization. To address these limitations, we propose UGP, a pruned framework designed to enhance generalization power for LiDAR point cloud registration. The core insight in UGP is the elimination of cross-attention mechanisms to improve generalization, allowing the network to concentrate on intra-frame feature extraction. Additionally, we introduce a progressive self-attention module to reduce ambiguity in large-scale scenes and integrate Bird's Eye View (BEV) features to incorporate semantic information about scene elements. Together, these enhancements significantly boost the network's generalization performance. We validated our approach through various generalization experiments in multiple outdoor scenes. In cross-distance generalization experiments on KITTI and nuScenes, UGP achieved state-of-the-art mean Registration Recall rates of 94.5% and 91.4%, respectively. In cross-dataset generalization from nuScenes to KITTI, UGP achieved a state-of-the-art mean Registration Recall of 90.9%.
****

Structure-Aware Correspondence Learning for Relative Pose Estimation
Yihan Chen, Wenfei Yang, Huan Ren, Shifeng Zhang, Tianzhu Zhang, Feng Wu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11611-11621
Relative pose estimation provides a promising way for achieving object-agnostic pose estimation. Despite the success of existing 3D correspondence-based methods, the reliance on explicit feature matching suffers from small overlaps in visible regions and unreliable feature estimation for invisible regions. Inspired by humans' ability to assemble two object parts that have small or no overlapping regions by considering object structure, we propose a novel Structure-Aware Correspondence Learning method for Relative Pose Estimation, which consists of two key modules. First, a structure-aware keypoint extraction module is designed to locate a set of kepoints that can represent the structure of objects with different shapes and appearance, under the guidance of a keypoint based image reconstruction loss. Second, a structure-aware correspondence estimation module is designed to model the intra-image and inter-image relationships between keypoints to extract structure-aware features for correspondence estimation. By jointly leveraging these two modules, the proposed method can naturally estimate 3D-3D correspondences for unseen objects without explicit feature matching for precise relative pose estimation. Experimental results on the CO3D, Objaverse and LineMOD datasets demonstrate that the proposed method significantly outperforms prior methods, i.e., with 5.7^\circ reduction in mean angular error on the CO3D dataset.
****
LoRA Recycle: Unlocking Tuning-Free Few-Shot Adaptability in Visual Foundation Models by Recycling Pre-Tuned LoRAs
Zixuan Hu, Yongxian Wei, Li Shen, Chun Yuan, Dacheng Tao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25026-25037
Large Language Models (LLMs) such as ChatGPT demonstrate strong few-shot adaptability without requiring fine-tuning, positioning them ideal for data-limited and real-time applications. However, this adaptability has not yet been replicated in current Visual Foundation Models (VFMs), which require explicit fine-tuning with sufficient tuning data. Besides, the pretraining-finetuning paradigm has led to the surge of numerous task-specific modular components, such as Low-Rank Adaptation (LoRA). For the first time, we explore the potential of reusing diverse pre-tuned LoRAs without accessing their original training data, to achieve tuning-free few-shot adaptation in VFMs. Our framework, LoRA Recycle, distills a meta-LoRA from diverse pre-tuned LoRAs with a meta-learning objective, using synthetic data inversely generated from pre-tuned LoRAs themselves. The VFM, once equipped with the meta-LoRA, is empowered to solve new few-shot tasks in a single forward pass, akin to the in-context learning of LLMs. Additionally, we incorporate a double-efficient mechanism, accelerating the data-generation and meta-training process while maintaining or even improving performance. Extensive experiments across various few-shot classification benchmarks across both in- and cross-domain scenarios demonstrate the superiority of our framework. Code is available at https://github.com/Egg-Hu/LoRA-Recycle.
****
One2Any: One-Reference 6D Pose Estimation for Any Object
Mengya Liu, Siyuan Li, Ajad Chhatkuli, Prune Truong, Luc Van Gool, Federico Tombari; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6457-6467
6D object pose estimation remains challenging for many applications due to dependencies on complete 3D models, multi-view images, or training limited to specific object categories. These requirements make generalization to novel objects difficult for which neither 3D models nor multi-view images may be available. To address this, we propose a novel method One2Any that estimates the relative 6-degrees of freedom (DOF) object pose using only a single reference-single query RGB-D image, without prior knowledge of its 3D model, multi-view data, or category constraints. We treat object pose estimation as an encoding-decoding process: first, we obtain a comprehensive Reference Object Pose Embedding (ROPE) that encode

s an object's shape, orientation, and texture from a single reference view. Usin g this embedding, a U-Net-based pose decoding module produces Reference Object C oordinate (ROC) for new views, enabling fast and accurate pose estimation. This simple encoding-decoding framework allows our model to be trained on any pair-wi se pose data, enabling large-scale training and demonstrating great scalability. Experiments on multiple benchmark datasets demonstrate that our model generaliz es well to novel objects, achieving state-of-the-art accuracy and robustness eve n rivaling methods that require multi-view or CAD inputs, at a fraction of compu te.
****

PyTorchGeoNodes: Enabling Differentiable Shape Programs for 3D Shape Reconstruct ion

Sinisa Stekovic, Arslan Artykov, Stefan Ainetter, Mattia D'Urso, Friedrich Fraun dorfer; Proceedings of the Computer Vision and Pattern Recognition Conference (C VPR), 2025, pp. 16283-16292

We propose PyTorchGeoNodes, a differentiable module for reconstructing 3D object s and their parameters from images using interpretable shape programs. Unlike tr aditional CAD model retrieval, shape programs allow reasoning about semantic par ameters, editing, and a low memory footprint. Despite their potential, shape pro grams for 3D scene understanding have been largely overlooked. Our key contribut ion is enabling gradient-based optimization by parsing shape programs, or more p recisely procedural models designed in Blender, into efficient PyTorch code. Whi le there are many possible applications of our PyTochGeoNodes, we show that a co mbination of PyTorchGeoNodes with genetic algorithm is a method of choice to opt imize both discrete and continuous shape program parameters for 3D reconstructio n and understanding of 3D object parameters. Our modular framework can be furthe r integrated with other reconstruction algorithms, and we demonstrate one such i ntegration to enable procedural Gaussian splatting. Our experiments on the ScanN et dataset show that our method achieves accurate reconstructions while enabling , until now, unseen level of 3D scene understanding.
****

Contextual AD Narration with Interleaved Multimodal Sequence

Hanlin Wang, Zhan Tong, Kecheng Zheng, Yujun Shen, Limin Wang; Proceedings of th e Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8372-8383

The Audio Description (AD) task aims to generate descriptions of visual elements for visually impaired individuals to help them access long-form video contents, like movie. With video feature, text, character bank and context information as inputs, the generated ADs are able to correspond to the characters by name and provide reasonable, contextual descriptions to help audience understand the stor yline of movie. To achieve this goal, we propose to leverage pre-trained foundat ion models through a simple and unified framework to generate ADs with interleav ed multimodal sequence as input, termed as Uni-AD. To enhance the alignment of f eatures across various modalities with finer granularity, we introduce a simple and lightweight module that maps video features into the textual feature space. Moreover, we also propose a character-refinement module to provide more precise information by identifying the main characters who play more significant role in the video context. With these unique designs, we further incorporate contextual information and a contrastive loss into our architecture to generate more smoot h and contextual ADs. Experiments on the MAD-eval dataset show that Uni-AD can a chieve state-of-the-art performance on AD generation, which demonstrates the eff ectiveness of our approach. Our code is available at: https://github.com/ant-res earch/UniAD.
****

FIFA: Fine-grained Inter-frame Attention for Driver's Video Gaze Estimation

Daosong Hu, Mingyue Cui, Kai Huang; Proceedings of the Computer Vision and Patte rn Recognition Conference (CVPR), 2025, pp. 18760-18769

Gaze direction serves as a pivotal indicator for assessing the level of driver a ttention. While image-based gaze estimation has been extensively researched, the re has been a recent shift towards capturing gaze direction from video sequences . This approach encounters notable challenges, including the comprehension of th

e dynamic pupil evolution across frames and the extraction of head pose informat ion from a relatively static background. To surmount these challenges, we introd uce a dual-stream deep learning framework that explicitly models the displacemen t changes of the pupil through a fine-grained inter-frame attention mechanism an d generates weights to adjust gaze embeddings. This technique transforms the fac e into a set of distinct patches and employs cross-attention to ascertain the co rrelation between pixel displacements in various patches and adjacent frames, th ereby tracking spatial dynamics within the sequence. Our method is validated usi ng two publicly available driver gaze datasets, and the results indicate that it achieves state-of-the-art performance or is on par with the best outcomes while reducing the parameters.

****

MNE-SLAM: Multi-Agent Neural SLAM for Mobile Robots

Tianchen Deng, Guole Shen, Chen Xun, Shenghai Yuan, Tongxin Jin, Hongming Shen, Yanbo Wang, Jingchuan Wang, Hesheng Wang, Danwei Wang, Weidong Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1485-1494

Neural implicit scene representations have recently shown promising results in d ense visual SLAM. However, existing implicit SLAM algorithms are constrained to single-agent scenarios, and fall difficulty in large indoor scenes and long sequ ences. Existing multi-agent SLAM frameworks cannot meet the constraints of commu nication bandwidth. To this end, we propose the first distributed multi-agent co llaborative SLAM framework with distributed mapping and camera tracking, joint s cene representation, intra-to-inter loop closure, and multi-submap fusion. Speci fically, our proposed distributed neural mapping and tracking framework only nee ds peer-to-peer communication, which can greatly improve multi-agent cooperation and communication efficiency. A novel intra-to-inter loop closure method is des igned to achieve local (single-agent) and global (multi-agent) consistency. Furt hermore, to the best of our knowledge, there is no real-world dataset for NeRF-b ased/GS-based SLAM that provides both continuous-time trajectories groundtruth a nd high-accuracy 3D meshes groundtruth. To this end, we propose the first real-w orld indoor neural slam (INS) dataset covering both single-agent and multi-agent scenarios, ranging from small room to large-scale scenes, with high-accuracy gr ound truth for both 3D mesh and continuous-time camera trajectory. This dataset can advance the development of the community. Experiments on various datasets de monstrate the superiority of the proposed method in both mapping, tracking, and communication. The dataset and code will be open-source on https://github.com/dt c111111/MNESLAM.

****

TensoFlow: Tensorial Flow-based Sampler for Inverse Rendering

Chun Gu, Xiaofei Wei, Li Zhang, Xiatian Zhu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 495-504

Inverse rendering aims to recover scene geometry, material properties, and light ing from multi-view images. Given the complexity of light-surface interactions, importance sampling is essential for the evaluation of the rendering equation, a s it reduces variance and enhances the efficiency of Monte Carlo sampling. Exist ing inverse rendering methods typically use pre-defined non-learnable importance samplers in prior manually, struggling to effectively match the spatially and d irectionally varied integrand and resulting in high variance and suboptimal perf ormance. To address this limitation, we propose the concept of learning a spatia lly and directionally aware importance sampler for the rendering equation to acc urately and flexibly capture the unconstrained complexity of a typical scene. We further formulate TensoFlow, a generic approach for sampler learning in inverse rendering, enabling to closely match the integrand of the rendering equation sp atially and directionally. Concretely, our sampler is parameterized by normalizi ng flows, allowing both directional sampling of incident light and probability d ensity function (PDF) inference. To capture the characteristics of the sampler s patially, we learn a tensorial representation of the scene space, which imposes spatial conditions, together with reflected direction, leading to spatially and directionally aware sampling distributions. Our model can be optimized by minimi

zing the difference between the integrand and our normalizing flow. Extensive ex periments validate the superiority of TensoFlow over prior alternatives on both synthetic and real-world benchmarks.
****

FRAMES-VQA: Benchmarking Fine-Tuning Robustness across Multi-Modal Shifts in Vis ual Question Answering
Chengyue Huang, Brisa Maneechotesuwan, Shivang Chopra, Zsolt Kira; Proceedings o f the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3909-3918
Visual question answering (VQA) systems face significant challenges when adaptin g to real-world data shifts, especially in multi-modal contexts. While robust fi ne-tuning strategies are essential for maintaining performance across in-distrib ution (ID) and out-of-distribution (OOD) scenarios, current evaluation settings are primarily unimodal or particular to some types of OOD, offering limited ins ight into the complexities of multi-modal contexts. In this work, we propose a n ew benchmark FRAMES-VQA (Fine-Tuning Robustness Across Multi-Modal Shifts in VQA) for evaluating robust fine-tuning for VQA tasks. We utilize ten existing VQA b enchmarks, including VQAv2, IV-VQA, VQA-CP, OK-VQA and others, and categorize th em into ID, near and far OOD datasets covering uni-modal, multi-modal and advers arial distribution shifts. We first conduct a comprehensive comparison of existi ng robust fine-tuning methods. We then quantify the distribution shifts by calcu lating the Mahalanobis distance using uni-modal and multi-modal embeddings extra cted from various models. Further, we perform an extensive analysis to explore t he interactions between uni- and multi-modal shifts as well as modality importan ce for ID and OOD samples. These analyses offer valuable guidance on developing more robust fine-tuning methods to handle multi-modal distribution shifts. The c ode is available at https://github.com/chengyuehuang511/FRAMES-VQA.
****

Shape Abstraction via Marching Differentiable Support Functions
Sunkyung Park, Jeongmin Lee, Dongjun Lee; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16902-16911
Shape abstraction, simplifying shape representation into a set of primitives, is a fundamental topic in computer vision. The choice of primitives shapes the str ucture of world understanding, yet achieving both high abstraction accuracy and versatility remains challenging. In this paper, we introduce a novel framework f or shape abstraction utilizing a differentiable support function (DSF), which of fers unique advantages in representing a wide range of convex shapes with fewer parameters, providing smooth surface approximation and enabling differentiable c ontact features (gap, point, normal) essential for downstream applications invol ving contact-related problems. To tackle the associated optimization and combina torial challenges, we introduce two techniques: differentiable shape parameteriz ation and hyperplane-based marching to enhance accuracy and reduce DSF requireme nts. We validate our method through experiments demonstrating superior accuracy and efficiency, and showcase its applicability in tasks requiring differentiable contact information.
****

LSceneLLM: Enhancing Large 3D Scene Understanding Using Adaptive Visual Preferen ces
Hongyan Zhi, Peihao Chen, Junyan Li, Shuailei Ma, Xinyu Sun, Tianhang Xiang, Yin jie Lei, Mingkui Tan, Chuang Gan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3761-3771
Research on 3D Vision-Language Models (3D-VLMs) is gaining increasing attention, which is crucial for developing embodied AI within 3D scenes, such as visual na vigation and embodied question answering. Due to the high density of visual feat ures, especially in large 3D scenes, accurately locating task-relevant visual in formation is challenging. Existing works attempt to segment all objectsand consi der their features as scene representations. However, these task-agnostic object features include much redundant information and missing details for the task-re levant area. To tackle these problems, we propose LSceneLLM, an adaptive framewo rk that automatically identifies task-relevant areas by leveraging LLM's visual

preference for different tasks, followed by a plug-and-play scene magnifier module to capture fine-grained details in focused areas. Specifically, a dense token selector examines the attention map of LLM to identify visual preferences for the instruction input. It then magnifies fine-grained details of the focusing area. An adaptive self-attention module is leveraged to fuse the coarse-grained and selected fine-grained visual information. To comprehensively evaluate the large scene understanding ability of 3D-VLMs, we further introduce a cross-room understanding benchmark, XR-Scene, which contains a series of large scene understanding tasks including XR-QA, XR-EmbodiedPlanning, and XR-SceneCaption. Experiments show that our method surpasses existing methods on both large scene understanding and existing scene understanding benchmarks. Plunging our scene magnifier module into the existing 3D-VLMs also brings significant improvement.
****

Event Fields: Capturing Light Fields at High Speed, Resolution, and Dynamic Range

Ziyuan Qu, Zihao Zou, Vivek Boominathan, Praneeth Chakravarthula, Adithya Pediredla; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26910-26920

Event cameras, which feature pixels that independently respond to changes in brightness, are becoming increasingly popular in high-speed applications due to their lower latency, reduced bandwidth requirements, and enhanced dynamic range compared to traditional frame-based cameras. Numerous imaging and vision techniques have leveraged event cameras for high-speed scene understanding by capturing high-framerate, high-dynamic range videos, primarily utilizing the temporal advantages inherent to event cameras. Additionally, imaging and vision techniques have utilized the light field---a complementary dimension to temporal information---for enhanced scene understanding. In this work, we propose "Event Fields", a new approach that utilizes innovative optical designs for event cameras to capture light fields at high speed. We develop the underlying mathematical framework for Event Fields and introduce two foundational frameworks to capture them practically: spatial multiplexing to capture temporal derivatives and temporal multiplexing to capture angular derivatives. To realize these, we design two complementary optical setups---one using a kaleidoscope for spatial multiplexing and another using a galvanometer for temporal multiplexing. We evaluate the performance of both designs using a custom-built simulator and real hardware prototypes, showcasing their distinct benefits. Our event fields unlock the full advantages of typical light fields--like post-capture refocusing and depth estimation--now supercharged for high-speed and high-dynamic range scenes. This novel light-sensing paradigm opens doors to new applications in photography, robotics, and AR/VR, and presents fresh challenges in rendering and machine learning.
****

HyperFree: A Channel-adaptive and Tuning-free Foundation Model for Hyperspectral Remote Sensing Imagery

Jingtao Li, Yingyi Liu, Xinyu Wang, Yunning Peng, Chen Sun, Shaoyu Wang, Zhendong Sun, Tian Ke, Xiao Jiang, Tangwei Lu, Anran Zhao, Yanfei Zhong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23048-23058

Advanced interpretation of hyperspectral remote sensing images benefits many precise Earth observation tasks. Recently, visual foundation models have promoted the remote sensing interpretation but concentrating on RGB and multispectral images. Due to the varied hyperspectral channels, existing foundation models would face image-by-image tuning situation, imposing great pressure on hardware and time resources. In this paper, we propose a tuning-free hyperspectral foundation model called HyperFree, by adapting the existing visual prompt engineering. To process varied channel numbers, we design a learned weight dictionary covering full-spectrum from 0.4 2.5um, supporting to build the embedding layer dynamically. To make the prompt design more tractable, HyperFree can generate multiple semantic-aware masks for one prompt by treating feature distance as semantic-similarity. After pre-training HyperFree on constructed large-scale high-resolution hyperspectral images, HyperFree (1 prompt) has shown comparable results with specializ

ed models (5 shots) on 5 tasks and 11 datasets. Code and dataset would be access ible at https://rsidea.whu.edu.cn/hyperfree.htm#.
****

Exploring Temporally-Aware Features for Point Tracking
Inès Hyeonsu Kim, Seokju Cho, Jiahui Huang, Jung Yi, Joon-Young Lee, Seungryong Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1962-1972
Point tracking in videos is a fundamental task with applications in robotics, vi deo editing, and more. While many vision tasks benefit from pre-trained feature backbones to improve generalizability, point tracking has primarily relied on si mpler backbones trained from scratch on synthetic data, which may limit robustne ss in real-world scenarios. Additionally, point tracking requires temporal aware ness to ensure coherence across frames, but using temporally-aware features is s till underexplored. Most current methods often employ a two-stage process: an in itial coarse prediction followed by a refinement stage to inject temporal inform ation and correct errors from the coarse stage. These approach, however, is comp utationally expensive and potentially redundant if the feature backbone itself c aptures sufficient temporal information. In this work, we introduce Chrono, a fe ature backbone specifically designed for point tracking with built-in temporal a wareness. Leveraging pre-trained representations from self-supervised learner DI NOv2 and enhanced with a temporal adapter, Chrono effectively captures long-term temporal context, enabling precise prediction even without the refinement stage . Experimental results demonstrate that Chrono achieves state-of-the-art perform ance in a refiner-free setting on the TAP-Vid-DAVIS and TAP-Vid-Kinetics dataset s, among common feature backbones used in point tracking as well as DINOv2, with exceptional efficiency.
****

GBlobs: Explicit Local Structure via Gaussian Blobs for Improved Cross-Domain Li DAR-based 3D Object Detection
Dušan Mali■, Christian Fruhwirth-Reisinger, Samuel Schulter, Horst Possegger; Pr oceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025 , pp. 27357-27367
LiDAR-based 3D detectors need large datasets for training, yet they struggle to generalize to novel domains. Domain Generalization (DG) aims to mitigate this by training detectors that are invariant to such domain shifts. Current DG approac hes exclusively rely on global geometric features (point cloud Cartesian coordin ates) as input features. Over-reliance on these global geometric features can, h owever, cause 3D detectors to prioritize object location and absolute position, resulting in poor cross-domain performance. To mitigate this, we propose to expl oit explicit local point cloud structure for DG, in particular by encoding point cloud neighborhoods with Gaussian blobs, GBlobs. Our proposed formulation is hi ghly efficient and requires no additional parameters. Without any bells and whis tles, simply by integrating GBlobs in existing detectors, we beat the current st ate-of-the-art in challenging single-source DG benchmarks by over 21 mAP (Waymo- >KITTI), 13 mAP (KITTI->Waymo), and 12 mAP (nuScenes->KITTI), without sacrificin g in-domain performance. Additionally, GBlobs demonstrate exceptional performanc e in multi-source DG, surpassing the current state-of-the-art by 17, 12, and 5 m AP on Waymo, KITTI, and ONCE, respectively.
****

V^2Dial: Unification of Video and Visual Dialog via Multimodal Experts
Adnen Abdessaied, Anna Rohrbach, Marcus Rohrbach, Andreas Bulling; Proceedings o f the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8637- 8647
We present V2Dial - a novel expert-based model specifically geared towards simul taneously handling image and video input data for multimodal conversational task s. Current multimodal models primarily focus on simpler tasks (e.g., VQA, VideoQ A, video-text retrieval) and often neglect the more challenging conversational c ounterparts, such as video and visual/image dialog. Moreover, works on both conv ersational tasks evolved separately from each other despite their apparent simil arities limiting their applicability potential. To this end, we propose to unify

both tasks using a single model that for the first time jointly learns the spatial and temporal features of images and videos by routing them through dedicated experts and aligns them using matching and contrastive learning techniques. Furthermore, we systemically study the domain shift between the two tasks by investigating whether and to what extent these seemingly related tasks can mutually benefit from their respective training data. Extensive evaluations on the widely used video and visual dialog datasets of AVSD and VisDial show that our model achieves new state-of-the-art results across four benchmarks both in zero-shot and fine-tuning settings.

****

Detail-Preserving Latent Diffusion for Stable Shadow Removal

Jiamin Xu, Yuxin Zheng, Zelong Li, Chi Wang, Renshu Gu, Weiwei Xu, Gang Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7592-7602

Achieving high-quality shadow removal with strong generalizability is challenging in scenes with complex global illumination. Due to the limited diversity in shadow removal datasets, current methods are prone to overfitting training data, often leading to reduced performance on unseen cases. To address this, we leverage the rich visual priors of a pre-trained Stable Diffusion (SD) model and propose a two-stage fine-tuning pipeline to adapt the SD model for stable and efficient shadow removal. In the first stage, we fix the VAE and fine-tune the denoiser in latent space, which yields substantial shadow removal but may lose some high-frequency details. To resolve this, we introduce a second stage, called the detail injection stage. This stage selectively extracts features from the VAE encoder to modulate the decoder, injecting fine details into the final results. Experimental results show that our method outperforms state-of-the-art shadow removal techniques. The cross-dataset evaluation further demonstrates that our method generalizes effectively to unseen data, enhancing the applicability of shadow removal methods.

****

Scaling Down Text Encoders of Text-to-Image Diffusion Models

Lifu Wang, Daqing Liu, Xinchen Liu, Xiaodong He; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18424-18433

Text encoders in diffusion models have rapidly evolved, transitioning from CLIP to T5-XXL. Although this evolution has significantly enhanced the models' ability to understand complex prompts and generate text, it also leads to a substantial increase in the number of parameters. Despite T5 series encoders being trained on the C4 natural language corpus, which includes a significant amount of non-visual data, diffusion models with T5 encoder do not respond to those non-visual prompts, indicating redundancy in representational power. Therefore, it raises an important question: "Do we really need such a large text encoder?" In pursuit of an answer, we employ vision-based knowledge distillation to train a series of T5 encoder models. To fully inherit its capabilities, we constructed our dataset based on three criteria: image quality, semantic understanding, and text-rendering. Our results demonstrate the scaling down pattern that the distilled T5-base model can generate images of comparable quality to those produced by T5-XXL, while being 50 times smaller in size. This reduction in model size significantly lowers the GPU requirements for running state-of-the-art models such as FLUX and SD3, making high-quality text-to-image generation more accessible.

****

3D Gaussian Head Avatars with Expressive Dynamic Appearances by Compact Tensorial Representations

Yating Wang, Xuan Wang, Ran Yi, Yanbo Fan, Jichen Hu, Jingcheng Zhu, Lizhuang Ma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21117-21126

Recent studies have combined 3D Gaussian and 3D Morphable Models (3DMM) to construct high-quality 3D head avatars. In this line of research, existing methods either fail to capture the dynamic textures or incur significant overhead in terms of runtime speed or storage space. To this end, we propose a novel method that addresses all the aforementioned demands. In specific, we introduce an expressiv

e and compact representation that encodes texture-related attributes of the 3D G
aussians in the tensorial format. We store appearance of neutral expression in s
tatic tri-planes, and represents dynamic texture details for different expressio
ns using lightweight 1D feature lines, which are then decoded into opacity offse
t relative to the neutral face. We further propose adaptive truncated opacity pe
nalty and class-balanced sampling to improve generalization across different exp
ressions. Experiments show this design enables accurate face dynamic details cap
turing while maintains real-time rendering and significantly reduces storage cos
ts, thus broadening the applicability to more scenarios.
****

MambaIRv2: Attentive State Space Restoration
Hang Guo, Yong Guo, Yaohua Zha, Yulun Zhang, Wenbo Li, Tao Dai, Shu-Tao Xia, Yaw
ei Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CV
PR), 2025, pp. 28124-28133
The Mamba-based image restoration backbones have recently demonstrated significa
nt potential in balancing global reception and computational efficiency. However
, the inherent causal modeling limitation of Mamba, where each token depends sol
ely on its predecessors in the scanned sequence, restricts the full utilization
of pixels across the image and thus presents new challenges in image restoration
. In this work, we propose MambaIRv2, which equips Mamba with the non-causal mod
eling ability similar to ViTs to reach the attentive state space restoration mod
el. Specifically, the proposed attentive state-space equation allows to attend b
eyond the scanned sequence and facilitate image unfolding with just one single s
can. Moreover, we further introduce a semantic-guided neighboring mechanism to e
ncourage interaction between distant but similar pixels. Extensive experiments s
how our MambaIRv2 outperforms SRFormer by even 0.35dB PSNR for lightweight SR ev
en with 9.3% less parameters and suppresses HAT on classic SR by up to 0.29dB.
****

Floating No More: Object-Ground Reconstruction from a Single Image
Yunze Man, Yichen Sheng, Jianming Zhang, Liang-Yan Gui, Yu-Xiong Wang; Proceedin
gs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2
7134-27143
Recent advancements in 3D object reconstruction from single images have primaril
y focused on improving the accuracy of object shapes. Yet, these techniques ofte
n fail to accurately capture the inter-relation between the object, ground, and
camera. As a result, the reconstructed objects often appear floating or tilted w
hen placed on flat surfaces. This limitation significantly affects 3D-aware imag
e editing applications like shadow rendering and object pose manipulation. To ad
dress this issue, we introduce ORG (Object Reconstruction with Ground), a novel
task aimed at reconstructing 3D object geometry in conjunction with the ground s
urface. Our method uses two compact pixel-level representations to depict the re
lationship between camera, object, and ground. Experiments show that the propose
d ORG model can effectively reconstruct object-ground geometry on unseen data, s
ignificantly enhancing the quality of shadow generation and pose manipulation co
mpared to conventional single-image 3D reconstruction techniques.
****

POT: Prototypical Optimal Transport for Weakly Supervised Semantic Segmentation
Jian Wang, Tianhong Dai, Bingfeng Zhang, Siyue Yu, Eng Gee Lim, Jimin Xiao; Proc
eedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025,
pp. 15055-15064
Weakly Supervised Semantic Segmentation (WSSS) leverages Class Activation Maps (
CAMs) to extract spatial information from image-level labels. However, CAMs prim
arily highlight the most discriminative foreground regions, leading to incomplet
e results. Prototype-based methods attempt to address this limitation by employi
ng prototype CAMs instead of classifier CAMs. Nevertheless, existing prototype-b
ased methods typically use a single prototype for each class, which is insuffici
ent to capture all attributes of the foreground features due to the significant
intra-class variations across different images. Consequently, these methods stil
l struggle with incomplete CAM predictions. In this paper, we propose a novel fr
amework called Prototypical Optimal Transport (POT) for WSSS. POT enhances CAM p

redictions by dividing features into multiple clusters and activating them separately using multiple cluster prototypes. In this process, a similarity-aware optimal transport is employed to assign features to the most probable clusters. This similarity-aware strategy ensures the prioritization of significant cluster prototypes, thereby improving the accuracy of feature assignment. Additionally, we introduce an adaptive OT-based consistency loss to refine feature representations. This framework effectively overcomes the limitations of single-prototype methods, providing more complete and accurate CAM predictions. Extensive experimental results on standard WSSS benchmarks (PASCAL VOC and MS COCO) demonstrate that our method significantly improves the quality of CAMs and achieves state-of-the-art performances. The source code will be released here https://github.com/jianwang91/POT.

****

CrossOver: 3D Scene Cross-Modal Alignment

Sayan Deb Sarkar, Ondrej Miksik, Marc Pollefeys, Daniel Barath, Iro Armeni; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8985-8994

Multi-modal 3D object understanding has gained significant attention, yet current approaches often assume complete data availability and rigid alignment across all modalities. We present CrossOver, a novel framework for cross-modal 3D scene understanding via flexible, scene-level modality alignment. Unlike traditional methods that require aligned modality data for every object instance, CrossOver learns a unified, modality-agnostic embedding space for scenes by aligning modalities -- RGB images, point clouds, CAD models, floorplans, and text descriptions -- with relaxed constraints and without explicit object semantics. Leveraging dimensionality-specific encoders, a multi-stage training pipeline, and emergent cross-modal behaviors, CrossOver supports robust scene retrieval and object localization, even with missing modalities. Evaluations on ScanNet and 3RScan datasets show its superior performance across diverse metrics, highlighting CrossOver's adaptability for real-world applications in 3D scene understanding.

****

Rethinking Temporal Fusion with a Unified Gradient Descent View for 3D Semantic Occupancy Prediction

Dubing Chen, Huan Zheng, Jin Fang, Xingping Dong, Xianfei Li, Wenlong Liao, Tao He, Pai Peng, Jianbing Shen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1505-1515

We present GDFusion, a temporal fusion method for vision-based 3D semantic occupancy prediction (VisionOcc). GDFusion opens up the underexplored aspects of temporal fusion within the VisionOcc framework, with a focus on both temporal cues and fusion strategies. It systematically examines the entire VisionOcc pipeline, identifying three fundamental yet previously overlooked temporal cues: scene-level consistencies, motion calibration, and geometric complementation. These cues capture diverse facets of temporal evolution and provide distinctive contributions across various modules in the general VisionOcc framework. To effectively fuse temporal signals across heterogeneous representations, we propose a novel fusion strategy by reinterpreting the formulation of vanilla RNNs. This reinterpretation leverages gradient descent on features to unify the integration of diverse temporal information, seamlessly embedding the proposed temporal cues into the network. Extensive experiments on nuScenes demonstrate that GDFusion significantly outperforms established baselines, achieving 2.2%--4.7% mIoU improvement while reducing memory consumption by 30%--72%. Codes are available at https://github.com/cdb342/GDFusion.

****

SKE-Layout: Spatial Knowledge Enhanced Layout Generation with LLMs

Junsheng Wang, Nieqing Cao, Yan Ding, Mengying Xie, Fuqiang Gu, Chao Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19414-19423

Generating layouts from textual descriptions by large language models (LLMs) plays a crucial role in precise spatial reasoning-induced domains such as robotic object rearrangement and text-to-image generation. However, current methods face

challenges in limited real-world examples, handling diverse layout descriptions and varying levels of granularity. To address these issues, a novel framework named Spatial Knowledge Enhanced Layout (SKE-Layout), is introduced. SKE-Layout integrates mixed spatial knowledge sources, leveraging both real and synthetic data to enhance spatial contexts. It utilizes diverse representations tailored to specific tasks and employs contrastive learning and multitask learning techniques for accurate spatial knowledge retrieval. This framework generates more accurate and fine-grained visual layouts for object rearrangement and text-to-image generation tasks, achieving improvements of 5%-30% compared to existing methods.
****

Gaussian Eigen Models for Human Heads
Wojciech Zielonka, Timo Bolkart, Thabo Beeler, Justus Thies; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15930-15940
Current personalized neural head avatars face a trade-off: lightweight models lack detail and realism, while high-quality, animatable avatars require significant computational resources, making them unsuitable for commodity devices. To address this gap, we introduce Gaussian Eigen Models (GEM), which provide high-quality, lightweight, and easily controllable head avatars. GEM utilizes 3D Gaussian primitives for representing the appearance combined with Gaussian splatting for rendering. Building on the success of mesh-based 3D morphable face models (3DMM), we define GEM as an ensemble of linear eigenbases for representing the head appearance of a specific subject. In particular, we construct linear bases to represent the position, scale, rotation, and opacity of the 3D Gaussians. This allows us to efficiently generate Gaussian primitives of a specific head shape by a linear combination of the basis vectors, only requiring a low-dimensional parameter vector that contains the respective coefficients. We propose to construct these linear bases (GEM) by distilling high-quality compute-intense CNN-based Gaussian avatar models that can generate expression-dependent appearance changes like wrinkles. These high-quality models are trained on multi-view videos of a subject and are distilled using a series of principle component analyses. Once we have obtained the bases that represent the animatable appearance space of a specific human, we learn a regressor that takes a single RGB image as input and predicts the low-dimensional parameter vector that corresponds to the shown facial expression. We demonstrate that this regressor can be trained such that it effectively supports self- and cross-person reenactment from monocular videos without requiring prior mesh-based tracking. In a series of experiments, we compare GEM's self-reenactment and cross-person reenactment results to state-of-the-art 3D avatar methods, demonstrating GEM's higher visual quality and better generalization to new expressions. As our distilled linear model is highly efficient in generating novel animation states, we also show a real-time demo of GEMs driven by monocular webcam videos.
****

Scalable Video-to-Dataset Generation for Cross-Platform Mobile Agents
Yunseok Jang, Yeda Song, Sungryull Sohn, Lajanugen Logeswaran, Tiange Luo, Dong-Ki Kim, Kyunghoon Bae, Honglak Lee; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8604-8614
Recent advancements in Large Language Models (LLMs) and Vision-Language Models (VLMs) have sparked significant interest in developing GUI visual agents. We introduce MONDAY (Mobile OS Navigation Task Dataset for Agents from YouTube), a large-scale dataset of 313K annotated frames from 20K instructional videos capturing diverse real-world mobile OS navigation across multiple platforms. Models that include MONDAY in their pretraining phases demonstrate robust cross-platform generalization capabilities, consistently outperforming models trained on existing single OS datasets while achieving an average performance gain of 18.11%p on an unseen mobile OS platform. To enable continuous dataset expansion as mobile platforms evolve, we present an automated framework that leverages publicly available video content to create comprehensive task datasets without manual annotation. Our framework comprises robust OCR-based scene detection (95.04% F1-score), near-perfect UI element detection (99.87% hit ratio), and novel multi-step action identification to extract reliable action sequences across diverse interface conf

igurations. We contribute both the MONDAY dataset and our automated collection f ramework to facilitate future research in mobile OS navigation, available at htt ps://github.com/runamu/monday.
****

Pattern Analogies: Learning to Perform Programmatic Image Edits by Analogy
Aditya Ganeshan, Thibault Groueix, Paul Guerrero, Radomir Mech, Matthew Fisher, Daniel Ritchie; Proceedings of the Computer Vision and Pattern Recognition Confe rence (CVPR), 2025, pp. 28715-28725
Pattern images are everywhere in the digital and physical worlds, and tools to e dit them are valuable. But editing pattern images is tricky: desired edits are o ften *programmatic*: structure-aware edits that alter the underlying program whi ch generates the pattern. One could attempt to infer this underlying program, bu t current methods for doing so struggle with complex images and produce unorgani zed programs that make editing tedious. In this work, we introduce a novel appro ach to perform programmatic edits on pattern images. By using a *pattern analogy *—a pair of simple patterns to demonstrate the intended edit—and a l earning-based generative model to execute these edits, our method allows users t o intuitively edit patterns. To enable this paradigm, we introduce **SplitWeave* *, a domain-specific language that, combined with a framework for sampling synth etic pattern analogies, enables the creation of a large, high-quality synthetic training dataset. We also present **TriFuser**, a Latent Diffusion Model (LDM) d esigned to overcome critical issues that arise when naively deploying LDMs to th is task. Extensive experiments on real-world, artist-sourced patterns reveals th at our method faithfully performs the demonstrated edit while also generalizing to related pattern styles beyond its training distribution.
****

4D-Fly: Fast 4D Reconstruction from a Single Monocular Video
Diankun Wu, Fangfu Liu, Yi-Hsin Hung, Yue Qian, Xiaohang Zhan, Yueqi Duan; Proce edings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, p p. 16663-16673
4D reconstruction from a single monocular video is an important but challenging task due to its inherent under-constrained nature. While most existing 4D recons truction methods focus on multi-camera settings, they always suffer from limited multi-view information in monocular videos. Recent studies have attempted to mi tigate the ill-posed problem by incorporating data-driven priors as additional s upervision. However, they require hours of optimization to align the splatted 2D feature maps of explicit Gaussians with various priors, which limits the range of applications. To address the time-consuming issue, we propose 4D-Fly, an effi cient and effective framework for reconstructing the 4D scene from a monocular v ideo (hundreds of frames within 6 minutes), more than 20 xfaster and even achiev ing higher quality than previous optimization methods. Our key insight is to unl eash the explicit property of Gaussian primitives and directly apply data priors to them. Specifically, we build a streaming 4D reconstruction paradigm that inc ludes: propagating existing Gaussian to the next timestep with an anchor-based s trategy, expanding the 4D scene map with the canonical Gaussian map, and an effi cient 4D scene optimization process to further improve visual quality and motion accuracy. Extensive experiments demonstrate the superiority of our 4D-Fly over state-of-the-art methods in terms of speed and quality.
****

STAR-Edge: Structure-aware Local Spherical Curve Representation for Thin-walled Edge Extraction from Unstructured Point Clouds
Zikuan Li, Honghua Chen, Yuecheng Wang, Sibo Wu, Mingqiang Wei, Jun Wang; Procee dings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp . 27254-27263
Extracting geometric edges from unstructured point clouds remains a significant challenge, particularly in thin-walled structures that are commonly found in eve ryday objects. Traditional geometric methods and recent learning-based approache s frequently struggle with these structures, as both rely heavily on sufficient contextual information from local point neighborhoods. However, 3D measurement d ata of thin-walled structures often lack the accurate, dense, and regular neighb

orhood sampling required for reliable edge extraction, resulting in degraded performance. In this work, we introduce STAR-Edge, a novel approach designed for detecting and refining edge points in thin-walled structures. Our method leverages a unique representation, the local spherical curve, to create structure-aware neighborhoods that emphasize co-planar points while reducing interference from close-by, non-co-planar surfaces. This representation is transformed into a rotation-invariant descriptor, which, combined with a lightweight multi-layer perceptron, enables robust edge point classification even in the presence of noise and sparse or irregular sampling. Besides, we also use the local spherical curve representation to estimate more precise normals and introduce an optimization function to project initially identified edge points exactly on the true edges. Experiments conducted on the ABC dataset and thin-walled structure-specific datasets demonstrate that STAR-Edge outperforms existing edge detection methods, showcasing better robustness under various challenging conditions. The source code is available at https://github.com/Miraclelzk/STAR-Edge.

****

Tokenize Image Patches: Global Context Fusion for Effective Haze Removal in Large Images

Jiuchen Chen, Xinyu Yan, Qizhi Xu, Kaiqi Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2258-2268

Global contextual information and local detail features are essential for haze removal tasks. Deep learning models perform well on small, low-resolution images, but they encounter difficulties with large, high-resolution ones due to GPU memory limitations. As a compromise, they often resort to image slicing or downsampling. The former diminishes global information, while the latter discards high-frequency details. To address these challenges, we propose DehazeXL, a haze removal method that effectively balances global context and local feature extraction, enabling end-to-end modeling of large images on mainstream GPU hardware. Additionally, to evaluate the efficiency of global context utilization in haze removal performance, we design a visual attribution method tailored to the characteristics of haze removal tasks. Finally, recognizing the lack of benchmark datasets for haze removal in large images, we have developed an ultra-high-resolution haze removal dataset (8KDehaze) to support model training and testing. It includes 10000 pairs of clear and hazy remote sensing images, each sized at 8192 x8192 pixels. Extensive experiments demonstrate that DehazeXL can infer images up to 10240 x10240 pixels with only 21 GB of memory, achieving state-of-the-art results among all evaluated methods. The source code and experimental dataset will soon be made publicly available.

****

Complementary Advantages: Exploiting Cross-Field Frequency Correlation for NIR-Assisted Image Denoising

Yuchen Wang, Hongyuan Wang, Lizhi Wang, Xin Wang, Lin Zhu, Wanxuan Lu, Hua Huang ; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12679-12689

Existing single-image denoising algorithms often struggle to restore details when dealing with complex noisy images. The introduction of near-infrared (NIR) images offers new possibilities for RGB image denoising. However, due to the inconsistency between NIR and RGB images, the existing works still struggle to balance the contributions of two fields in the process of image fusion. In response to this, in this paper, we develop a cross-field Frequency Correlation Exploiting Network (FCENet) for NIR-assisted image denoising. We first propose the frequency correlation prior based on an in-depth statistical frequency analysis of NIR-RGB image pairs. The prior reveals the complementary correlation of NIR and RGB images in the frequency domain. Leveraging frequency correlation prior, we then establish a frequency learning framework composed of Frequency Dynamic Selection Mechanism (FDSM) and Frequency Exhaustive Fusion Mechanism (FEFM). FDSM dynamically selects complementary information from NIR and RGB images in the frequency domain, and FEFM strengthens the control of common and differential features during the fusion process of NIR and RGB features. Extensive experiments on simulated and real data validate that the proposed method outperforms other state-of-the-

art methods. The code will be released at https://github.com/yuchenwang815/FCENet.

****

Eval3D: Interpretable and Fine-grained Evaluation for 3D Generation

Shivam Duggal, Yushi Hu, Oscar Michel, Aniruddha Kembhavi, William T. Freeman, Noah A. Smith, Ranjay Krishna, Antonio Torralba, Ali Farhadi, Wei-Chiu Ma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13326-13336

Despite the unprecedented progress in the field of 3D generation, current systems still often fail to produce high-quality 3D assets that are visually appealing and geometrically and semantically consistent across multiple viewpoints. To effectively assess the quality of the generated 3D data, there is a need for a reliable 3D evaluation tool. Unfortunately, existing 3D evaluation metrics often overlook the geometric quality of generated assets or merely rely on black-box multimodal large language models for coarse assessment. In this paper, we introduce Eval3D, a fine-grained, interpretable evaluation tool that can faithfully evaluate the quality of generated 3D assets based on various distinct yet complementary criteria. Our key observation is that many desired properties of 3D generation, such as semantic and geometric consistency, can be effectively captured by measuring the consistency among various foundation models and tools. We thus leverage a diverse set of models and tools as probes to evaluate the inconsistency of generated 3D assets across different aspects. Compared to prior work, Eval3D provides pixel-wise measurement, enables accurate 3D spatial feedback, and aligns more closely with human judgments. We comprehensively evaluate existing 3D generation models using Eval3D and highlight the limitations and challenges of current models.

****

Boosting the Dual-Stream Architecture in Ultra-High Resolution Segmentation with Resolution-Biased Uncertainty Estimation

Rong Qin, Xingyu Liu, Jinglei Shi, Liang Lin, Jufeng Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25960-25970

Over the last decade, significant efforts have been dedicated to designing efficient models for the challenge of ultra-high resolution (UHR) semantic segmentation. These models mainly follow the dual-stream architecture and generally fall into three subcategories according to the improvement objectives, i.e., dual-stream ensemble, selective zoom, and complementary learning. However, most of them overly concentrate on crafting complex pipelines to pursue one of the above objectives separately, limiting the model performance in both accuracy and inference consumption. In this paper, we suggest simultaneously achieving these objectives by estimating resolution-biased uncertainties in low resolution stream. Here, the resolution-biased uncertainty refers to the degree of prediction unreliability primarily caused by resolution loss from down-sampling operations. Specifically, we propose a dual-stream UHR segmentation framework, where an estimator is used to assess resolution-biased uncertainties through the entropy map and high-frequency feature residual. The framework also includes a selector, an ensembler, and a complementer to boost the model with obtained estimations. They share the uncertainty estimations as the weights to choose difficult regions as the inputs for UHR stream, perform weighted fusion between distinct streams, and enhance the learning for important pixels, respectively. Experiment results demonstrate that our method achieves a satisfactory balance between accuracy and inference consumption against other state-of-the-art (SOTA) methods. The code is available in the https://github.com/Qinrong-NKU/RUE.

****

DiffLO: Semantic-Aware LiDAR Odometry with Diffusion-Based Refinement

Yongshu Huang, Chen Liu, Minghang Zhu, Sheng Ao, Chenglu Wen, Cheng Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17050-17059

LiDAR odometry is a critical module in autonomous driving systems, responsible for accurate localization by estimating the relative pose transformation between consecutive point cloud frames. However, existing studies frequently encounter c

hallenges with unreliable pose estimation, due to the lack of in-depth understan ding of scenario and the presence of noise interference. To address this challen ge, we propose DiffLO, a semantic-aware LiDAR odometry network with diffusion-ba sed refinement. To mitigate the impact of challenging cases such as dynamic, rep etitive patterns, and low textures, we introduce a semantic distillation method that integrates semantic information into the odometry task. This allows the net work to gain a semantic understanding of the scene, enabling it to focus more on the objects that are beneficial for pose estimation. Additionally, to enhance t he robustness, we propose a diffusion-based refinement method. This method uses pose-related features as conditional constraints for generative diversity, itera tively refining the pose estimation to achieve greater accuracy. Comparative exp eriments on the KITTI odometry dataset demonstrate that the proposed method achi eves state-of-the-art performance among existing learning-based approaches. Furt hermore, the proposed DiffLO method outperforms the classic A-LOAM on most evalu ation sequences.

****

pFedMxF: Personalized Federated Class-Incremental Learning with Mixture of Frequ ency Aggregation

Yifei Zhang, Hao Zhu, Alysa Ziying Tan, Dianzhi Yu, Longtao Huang, Han Yu; Proce edings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, p p. 30640-30650

Federated learning (FL) has emerged as a promising paradigm for privacy-preservi ng collaborative machine learning. However, extending FL to class incremental le arning settings introduces three key challenges: 1) spatial heterogeneity due to non-IID data distributions across clients, 2) temporal heterogeneity due to seq uential arrival of tasks, and 3) resource heterogeneity due to diverse client ca pabilities. Existing approaches generally address these challenges in isolation, potentially leading to interference between updates, catastrophic forgetting, o r excessive communication overhead. In this paper, we propose personalized Feder ated class-incremental parameter efficient fine-tuning with Mixture of Frequency aggregation (pFedMixF), a novel framework that simultaneously addresses all thr ee heterogeneity challenges through frequency domain decomposition. Our key insi ght is that assigning orthogonal frequency components to different clients and t asks enables interference-free learning to be achieved with minimal communicatio n costs. We further design an Auto-Task Agnostic Classifier that automatically r outes samples to task-specific classifiers while adapting to heterogeneous class distributions.We conduct extensive experiments on three benchmark datasets, com paring our approach with eight state-of-the-art methods. The results demonstrate that \methodname achieves comparable test accuracy while requiring only 25% o f the entire model parameters and incurring significantly lower communication co sts than baseline methods.

****

Style-Editor: Text-driven Object-centric Style Editing

Jihun Park, Jongmin Gim, Kyoungmin Lee, Seunghun Lee, Sunghoon Im; Proceedings o f the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18281 -18291

We present Text-driven object-centric style editing model named Style-Editor, a novel method that guides style editing at an object-centric level using textual inputs.The core of Style-Editor is our Patch-wise Co-Directional (PCD) loss, met iculously designed for precise object-centric editing that are closely aligned w ith the input text. This loss combines a patch directional loss for text-guided style direction and a patch distribution consistency loss for even CLIP embeddin g distribution across object regions. It ensures a seamless and harmonious style editing across object regions.Key to our method are the Text-Matched Patch Sele ction (TMPS) and Pre-fixed Region Selection (PRS) modules for identifying object locations via text, eliminating the need for segmentation masks. Lastly, we int roduce an Adaptive Background Preservation (ABP) loss to maintain the original s tyle and structural essence of the image's background. This loss is applied to d ynamically identified background areas.Extensive experiments underline the effec tiveness of our approach in creating visually coherent and textually aligned sty

le editing.
****
Transfer Your Perspective: Controllable 3D Generation from Any Viewpoint in a Dr
iving Scene
Tai-Yu Pan, Sooyoung Jeon, Mengdi Fan, Jinsu Yoo, Zhenyang Feng, Mark Campbell,
Kilian Q. Weinberger, Bharath Hariharan, Wei-Lun Chao; Proceedings of the Comput
er Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12027-12036
Self-driving cars relying solely on ego-centric perception face limitations in s
ensing, often failing to detect occluded, faraway objects. Collaborative autonom
ous driving (CAV) seems like a promising direction, but collecting data for deve
lopment is non-trivial. It requires placing multiple sensor-equipped agents in a
 real-world driving scene, simultaneously! As such, existing datasets are limite
d in locations and agents. We introduce a novel surrogate to the rescue, which i
s to generate realistic perception from different viewpoints in a driving scene,
 conditioned on a real-world sample -- the ego-car's sensory data. This surrogat
e has huge potential: it could potentially turn any ego-car dataset into a colla
borative driving one to scale up the development of CAV. We present the very fir
st solution, using a combination of synthetic collaborative data and real ego-ca
r data. Our method, Transfer Your Perspective (TYP), learns a conditioned diffus
ion model whose output samples are not only realistic but also consistent in bot
h semantics and layouts with the given ego-car data. Empirical results demonstra
te TYP's effectiveness in aiding in a CAV setting. In particular, TYP enables us
 to (pre-)train collaborative perception algorithms like early and late fusion w
ith little or no real-world collaborative data, greatly facilitating downstream
CAV applications.
****
Efficient Transfer Learning for Video-language Foundation Models
Haoxing Chen, Zizheng Huang, Yan Hong, Yanshuo Wang, Zhongcai Lyu, Zhuoer Xu, Ju
n Lan, Zhangxuan Gu; Proceedings of the Computer Vision and Pattern Recognition
Conference (CVPR), 2025, pp. 29129-29138
Pre-trained vision-language models provide a robust foundation for efficient tra
nsfer learning across various downstream tasks. In the field of video action rec
ognition, mainstream approaches often introduce additional modules to capture te
mporal information. Although the additional modules increase the capacity of mod
el, enabling it to better capture video-specific inductive biases, existing meth
ods typically introduce a substantial number of new parameters and are prone to
catastrophic forgetting of previously acquired generalizable knowledge. In this
paper, we propose a parameter-efficient Multi-modal Spatio-Temporal Adapter (MST
A) to enhance the alignment between textual and visual representations, achievin
g a balance between generalizable knowledge and task-specific adaptation. Furthe
rmore, to mitigate over-fitting and enhance generalizability, we introduce a spa
tio-temporal description-guided consistency constraint. This constraint involves
 providing template inputs (e.g., "a video of \ cls\ ") to the trainable languag
e branch and LLM-generated spatio-temporal descriptions to the pre-trained langu
age branch, enforcing output consistency between the branches. This approach red
uces overfitting to downstream tasks and enhances the distinguishability of the
trainable branch within the spatio-temporal semantic space. We evaluate the effe
ctiveness of our approach across four tasks: zero-shot transfer, few-shot learni
ng, base-to-novel generalization, and fully-supervised learning. Compared to man
y state-of-the-art methods, our MSTA achieves outstanding performance across all
 evaluations, while using only 2-7% of the trainable parameters in the original
model.
****
Radio Frequency Ray Tracing with Neural Object Representation for Enhanced RF Mo
deling
Xingyu Chen, Zihao Feng, Kun Qian, Xinyu Zhang; Proceedings of the Computer Visi
on and Pattern Recognition Conference (CVPR), 2025, pp. 21339-21348
Radio frequency (RF) propagation modeling poses unique electromagnetic simulatio
n challenges. While recent neural representations have shown success in visible
spectrum rendering, the fundamentally different scales and physics of RF signals

require novel modeling paradigms. In this paper, we introduce RFScape, a novel framework that bridges the gap between neural scene representation and RF propagation modeling. Our key insight is that complex RF-object interactions can be captured through object-centric neural representations while preserving the composability of traditional ray tracing. Unlike previous approaches that either rely on crude geometric approximations or require dense spatial sampling of entire scenes, RFScape learns per-object electromagnetic properties and enables flexible scene composition. Through extensive evaluation on real-world RF testbeds, we demonstrate that our approach achieves 13 dB improvement over conventional ray tracing and 5 dB over state-of-the-art neural baselines in modeling accuracy, while requiring only sparse training samples.

****

ANNEXE: Unified Analyzing, Answering, and Pixel Grounding for Egocentric Interaction

Yuejiao Su, Yi Wang, Qiongyang Hu, Chuang Yang, Lap-Pui Chau; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9027-9038

Egocentric interaction perception is one of the essential branches in investigating human-environment interaction, which lays the basis for developing next-generation intelligent systems. However, existing egocentric interaction understanding methods cannot yield coherent textual and pixel-level responses simultaneously according to user queries, which lacks flexibility for varying downstream application requirements. To comprehend egocentric interactions exhaustively, this paper presents a novel task named Egocentric Interaction Reasoning and pixel Grounding (Ego-IRG). Taking an egocentric image with the query as input, Ego-IRG is the first task that aims to resolve the interactions through three crucial steps: analyzing, answering, and pixel grounding, which results in fluent textual and fine-grained pixel-level responses. Another challenge is that existing datasets cannot meet the conditions for the Ego-IRG task. To address this limitation, this paper creates the Ego-IRGBench dataset based on extensive manual efforts, which includes over 20k egocentric images with 1.6 million queries and corresponding multimodal responses about interactions. Moreover, we design a unified ANNEXE model to generate text- and pixel-level outputs utilizing multimodal large language models, which enables a comprehensive interpretation of egocentric interactions. The experiments on the Ego-IRGBench exhibit the effectiveness of our ANNEXE model compared with other works.

****

MET3R: Measuring Multi-View Consistency in Generated Images

Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, Jan Eric Lenssen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6034-6044

We introduce MEt3R, a metric for multi-view consistency in generated images. Large-scale generative models for multi-view image generation are rapidly advancing the field of 3D inference from sparse observations. However, due to the nature of generative modeling, traditional reconstruction metrics are not suitable to measure the quality of generated outputs and metrics that are independent of the sampling procedure are desperately needed. In this work, we specifically address the aspect of consistency between generated multi-view images, which can be evaluated independently of the specific scene. Our approach uses DUSt3R to obtain dense 3D reconstructions from image pairs in a feed-forward manner, which are used to warp image contents from one view into the other. Then, feature maps of these images are compared to obtain a similarity score that is invariant to view-dependent effects. Using MEt3R, we evaluate the consistency of a large set of previous methods for novel view and video generation, including our open, multi-view latent diffusion model. Code is available online: geometric-rl.mpi-inf.mpg.de/met3r/

****

Segmenting Maxillofacial Structures in CBCT Volumes

Federico Bolelli, Kevin Marchesini, Niels van Nistelrooij, Luca Lumetti, Vittorio Pipoli, Elisa Ficarra, Shankeeth Vinayahalingam, Costantino Grana; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 523

8-5248
Cone-beam computed tomography (CBCT) is a standard imaging modality in orofacial and dental practices, providing essential 3D volumetric imaging of anatomical structures, including jawbones, teeth, sinuses, and neurovascular canals. Accurately segmenting these structures is fundamental to numerous clinical applications, such as surgical planning and implant placement. However, manual segmentation of CBCT scans is time-intensive and requires expert input, creating a demand for automated solutions through deep learning. Effective development of such algorithms relies on access to large, well-annotated datasets, yet current datasets are often privately stored or limited in scope and considered structures, especially concerning 3D annotations. This paper proposes ToothFairy2, a comprehensive, publicly accessible CBCT dataset with voxel-level 3D annotations of 42 distinct classes corresponding to maxillofacial structures. We validate the dataset by benchmarking state-of-the-art neural network models, including convolutional, transformer-based, and hybrid Mamba-based architectures, to evaluate segmentation performance across complex anatomical regions. Our work also explores adaptations to the nnU-Net framework to optimize multi-class segmentation for maxillofacial anatomy. The proposed dataset provides a fundamental resource for advancing maxillofacial segmentation and supports future research in automated 3D image analysis in digital dentistry.
****

3D Dental Model Segmentation with Geometrical Boundary Preserving
Shufan Xi, Zexian Liu, Junlin Chang, Hongyu Wu, Xiaogang Wang, Aimin Hao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10476-10485
3D intraoral scan mesh is widely used in digital dentistry diagnosis, segmenting 3D intraoral scan mesh is a critical preliminary task. Numerous approaches have been devised for precise tooth segmentation. Currently, the deep learning-based methods are capable of the high accuracy segmentation of crown. However, the segmentation accuracy at the junction between the crown and the gum is still below average. Existing down-sampling methods are unable to effectively preserve the geometric details at the junction. To address these problems, we propose CrossTooth, a boundary-preserving segmentation method that combines 3D mesh selective downsampling to retain more vertices at the tooth-gingiva area, along with cross-modal discriminative boundary features extracted from multi-view rendered images, enhancing the geometric representation of the segmentation network. Using a point network as a backbone and incorporating image complementary features, CrossTooth significantly improves segmentation accuracy, as demonstrated by experiments on a public intraoral scan dataset.
****

Neuro-3D: Towards 3D Visual Decoding from EEG Signals
Zhanqiang Guo, Jiamin Wu, Yonghao Song, Jiahui Bu, Weijian Mai, Qihao Zheng, Wanli Ouyang, Chunfeng Song; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23870-23880
Human's perception of the visual world is shaped by the stereo processing of 3D information. Understanding how the brain perceives and processes 3D visual stimuli in the real world has been a longstanding endeavor in neuroscience. Towards this goal, we introduce a new neuroscience task: decoding 3D visual perception from EEG signals, a neuroimaging technique that enables real-time monitoring of neural dynamics enriched with complex visual cues. To provide the essential benchmark, we first present EEG-3D, a pioneering dataset featuring multimodal analysis data and extensive EEG recordings from 12 subjects viewing 72 categories of 3D objects rendered in both videos and images. Furthermore, we propose Neuro-3D, a 3D visual decoding framework based on EEG signals. This framework adaptively integrates EEG features derived from static and dynamic stimuli to learn complementary and robust neural representations, which are subsequently utilized to recover both the shape and color of 3D objects through the proposed diffusion-based colored point cloud decoder. To the best of our knowledge, we are the first to explore EEG-based 3D visual decoding. Experiments indicate that Neuro-3D not only reconstructs colored 3D objects with high fidelity, but also learns effective ne

ural representations that enable insightful brain region analysis. The code and dataset are available at https://github.com/gzq17/neuro-3D.
****

FastVLM: Efficient Vision Encoding for Vision Language Models
Pavan Kumar Anasosalu Vasu, Fartash Faghri, Chun-Liang Li, Cem Koc, Nate True, Albert Antony, Gokula Santhanam, James Gabriel, Peter Grasch, Oncel Tuzel, Hadi Pouransari; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19769-19780
Vision Language Models (VLMs) like LLaVA encode images into tokens aligned to the word embedding space of the LLM decoder. Scaling input image resolution is essential for improving performance, especially in text-rich image understanding tasks. However, popular visual encoders such as CLIP-pretrained ViTs become inefficient at high resolutions due to the large number of tokens and high encoding latency caused by stacked self-attention layers. At different operational resolutions, the vision encoder of a VLM can be optimized along two axes: reducing encoding latency and minimizing the number of visual tokens passed to the LLM, thereby lowering overall latency. In this work, we introduce FastVLM, which achieves an optimized trade-off between resolution, latency, and accuracy by incorporating FastViTHD--a new hybrid vision encoder that outputs fewer tokens and significantly reduces encoding time while processing high-resolution images. We provide a comprehensive efficiency analysis of the interplay between image resolution, vision latency, number of visual tokens, and LLM size. In the LLaVA-1.5 setup, we achieve 3.2x improvement in overall time-to-first-token (TTFT) while maintaining similar performance on VLM benchmarks compared to prior works. On text-rich evaluations like TextVQA and DocVQA, FastVLM obtains +8.4% and +12.5% better accuracy than ConvLLaVA at a similar operating point of 144 visual tokens. Compared to LLaVa-OneVision at the highest resolution (1152 x 1152), FastVLM achieves comparable performance on key benchmarks like SeedBench and MMMU, using the same LLM, but with 85x faster TTFT, 3x less vision instruction tuning data, and a vision encoder that is 3.4x smaller.
****

VISTA3D: A Unified Segmentation Foundation Model For 3D Medical Imaging
Yufan He, Pengfei Guo, Yucheng Tang, Andriy Myronenko, Vishwesh Nath, Ziyue Xu, Dong Yang, Can Zhao, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, Daguang Xu, Wenqi Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20863-20873
Foundation models for interactive segmentation in 2D natural images and videos have sparked significant interest in building 3D foundation models for medical imaging. However, the domain gaps and clinical use cases for 3D medical imaging require a dedicated model that diverges from existing 2D solutions. Specifically, such foundation models should support a full workflow that can actually reduce human effort. Treating 3D medical images as sequences of 2D slices and reusing interactive 2D foundation models seems straightforward, but 2D annotation is too time-consuming for 3D tasks. Moreover, for large cohort analysis, it's the highly accurate automatic segmentation models that reduce the most human effort. However, these models lack support for interactive corrections and lack zero-shot ability for novel structures, which is a key feature of "foundation". While reusing pre-trained 2D backbones in 3D enhances zero-shot potential, their performance on complex 3D structures still lags behind leading 3D models. To address these issues, we present VISTA3D, Versatile Imaging SegmenTation and Annotation model, that targets to solve all thesechallenges and requirements with one unified foundation model. VISTA3D is built on top of the well-established 3D segmentation pipeline, and it is the first model to achieve state-of-the-art performance in both 3D automatic (supporting 127 classes) and 3D interactive segmentation, even when compared with top 3D expert models on large and diverse benchmarks. Additionally, VISTA3D's 3D interactive design allows efficient human correction, and a novel 3D supervoxel method that distills 2D pretrained backbones grants VISTA3D top 3D zero-shot performance. We believe the model, recipe, and insights represent a promising step toward a clinically useful 3D foundation model. Code and weights are publicly available at https://github.com/Project-MONAI/VISTA

****
VideoGigaGAN: Towards Detail-rich Video Super-Resolution
Yiran Xu, Taesung Park, Richard Zhang, Yang Zhou, Eli Shechtman, Feng Liu, Jia-B in Huang, Difan Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2139-2149
Video super-resolution (VSR) models achieve temporal consistency but often produ ce blurrier results than their image-based counterparts due to limited generativ e capacity. This prompts the question: can we adapt a generative image upsampler for VSR while preserving temporal consistency? We introduce VideoGigaGAN, a new generative VSR model that combines high-frequency detail with temporal stabilit y, building on the large-scale GigaGAN image upsampler. Simple adaptations of Gi gaGAN for VSR led to flickering issues, so we propose techniques to enhance temp oral consistency. We validate the effectiveness of VideoGigaGAN by comparing it with state-of-the-art VSR models on public datasets and showcasing video results with 8x upsampling.
****
Probing the Mid-level Vision Capabilities of Self-Supervised Learning
Xuweiyi Chen, Markus Marks, Zezhou Cheng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30095-30105
Mid-level vision capabilities -- such as generic object localization and 3D geom etric understanding -- are not only fundamental to human vision but are also cru cial for many real-world applications of computer vision. These abilities emerge with minimal supervision during the early stages of human visual development. D espite their significance, current self-supervised learning (SSL) approaches are primarily designed and evaluated for high-level recognition tasks, leaving thei r mid-level vision capabilities largely unexamined. In this study, we introduce a suite of benchmark protocols to systematically assess mid-level vision capabil ities and present a comprehensive, controlled evaluation of 22 prominent SSL mod els across 8 mid-level vision tasks. Our experiments reveal a weak correlation b etween mid-level and high-level task performance. We also identify several SSL m ethods with highly imbalanced performance across mid-level and high-level capabi lities, as well as some that excel in both. Additionally, we investigate key fac tors contributing to mid-level vision performance, such as pretraining objective s and network architectures. Our study provides a holistic and timely view of wh at SSL models have learned, complementing existing research that primarily focus es on high-level vision tasks. We hope our findings guide future SSL research to benchmark models not only on high-level vision tasks but on mid-level as well.
****
S2D-LFE: Sparse-to-Dense Light Field Event Generation
Yutong Liu, Wenming Weng, Yueyi Zhang, Zhiwei Xiong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11207-11216
In this paper, we present S2D-LFE, an innovative approach for sparse-to-dense li ght field event generation. For the first time to our knowledge, S2D-LFE enables controllable novel view synthesis only from sparse-view light field event (LFE) data, and addresses three critical challenges for the LFE generation task: simp licity, controllability, and consistency. The simplicity aspect eliminates the d ependency on frame-based modality, which often suffers from motion blur and low frame-rate limitations. The controllability aspect enables precise view synthesi s under sparse LFE conditions with view-related constraints. The consistency asp ect ensures both cross-view and temporal coherence in the generated results. To realize S2D-LFE, we develop a novel diffusion-based generation network with two key components. First, we design an LFE-customized variational auto-encoder that effectively compresses and reconstructs LFE by integrating cross-view informati on. Second, we design an LFE-aware injection adaptor to extract comprehensive ge ometric and texture priors. Furthermore, we construct a large-scale synthetic LF E dataset containing 162 one-minute sequences using simulator, and capture a rea l-world testset using our custom-built sparse LFE acquisition system, covering d iverse indoor and outdoor scenes. Extensive experiments demonstrate that S2D-LFE successfully generates up to 9x9 dense LFE from sparse 2x2 inputs and outperfor ms existing methods on both synthetic and real-world data. The datasets and code

are available at https://github.com/Yutong2022/S2D-LFE.
****
The Art of Deception: Color Visual Illusions and Diffusion Models
Alexandra Gomez-Villa, Kai Wang, C.Alejandro Parraga, Bart■omiej Twardowski, Jesus Malo, Javier Vazquez-Corral, Joost van den Weijer; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18642-18652
Visual illusions in humans arise when interpreting out-of-distribution stimuli: if the observer is adapted to certain statistics, perception of outliers deviates from reality. Recent studies have shown that artificial neural networks (ANNs) can also be deceived by visual illusions.This revelation raises profound questions about the nature of visual information. Why are two independent systems, both human brains and ANNs, susceptible to the same illusions? Should any ANN be capable of perceiving visual illusions? Are these perceptions a feature or a flaw? In this work, we study how visual illusions are encoded in diffusion models. Remarkably, we show that they present human-like brightness/color shifts in their latent space. We use this fact to demonstrate that diffusion models can predict visual illusions. Furthermore, we also show how to generate new unseen visual illusions in realistic images using text-to-image diffusion models. We validate this ability through psychophysical experiments that show how our model-generated illusions also fool humans.
****
GLUS: Global-Local Reasoning Unified into A Single Large Language Model for Video Segmentation
Lang Lin, Xueyang Yu, Ziqi Pang, Yu-Xiong Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8658-8667
This paper proposes a novel framework utilizing multi-modal large language models (MLLMs) for referring video object segmentation (RefVOS). Previous MLLM-based methods commonly struggle with the dilemma between "Ref" and "VOS": they either specialize in understanding a few key frames (global reasoning) or tracking objects on continuous frames (local reasoning), and rely on external VOS or frame selectors to mitigate the other end of the challenge. However, our framework GLUS shows that Global and Local consistency can be Unified into a single video Segmentation MLLM: a set of sparse "context frames" provides global information, while a stream of continuous "query frames" conducts local object tracking. This is further supported by jointly training the MLLM with a pre-trained VOS memory bank to simultaneously digest short-range and long-range temporal information. To improve the information efficiency within the limited context window of MLLMs, we introduce object contrastive learning to distinguish hard false-positive objects and a self-refined framework to identify crucial frames and perform propagation. By collectively integrating these insights, our GLUS delivers a simple yet effective baseline, achieving new state-of-the-art for MLLMs on the MeViS and Ref-Youtube-VOS benchmark.
****
Progressive Rendering Distillation: Adapting Stable Diffusion for Instant Text-to-Mesh Generation without 3D Data
Zhiyuan Ma, Xinyue Liang, Rongyuan Wu, Xiangyu Zhu, Zhen Lei, Lei Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11036-11050
It is highly desirable to obtain a model that can generate high-quality 3D meshes from text prompts in just seconds. While recent attempts have adapted pre-trained text-to-image diffusion models, such as Stable Diffusion (SD), into generators of 3D representations (e.g., Triplane), they often suffer from poor quality due to the lack of sufficient high-quality 3D training data. Aiming at overcoming the data shortage, we propose a novel training scheme, termed as Progressive Rendering Distillation (PRD), eliminating the need for 3D ground-truths by distilling multi-view diffusion models and adapting SD into a native 3D generator. In each iteration of training, PRD uses the U-Net to progressively denoise the latent from random noise for a few steps, and in each step it decodes the denoised latent into 3D output. Multi-view diffusion models, including MVDream and RichDreamer, are used in joint with SD to distill text-consistent textures and geometrie

s into the 3D outputs through score distillation. Since PRD supports training wi thout 3D ground-truths, we can easily scale up the training data and improve gen eration quality for challenging text prompts with creative concepts. Meanwhile, PRD can accelerate the inference speed of the generation model in just a few ste ps. With PRD, we train a Triplane generator, namely TriplaneTurbo, which adds o nly 2.5% trainable parameters to adapt SD for Triplane generation. TriplaneTurbo outperforms previous text-to-3D generators in both efficiency and quality. Spe cifically, it can produce high-quality 3D meshes in 1.2 seconds and generalize w ell for challenging text input. The code is available at https://github.com/theE ricMa/TriplaneTurbo.
****

Efficient Long Video Tokenization via Coordinate-based Patch Reconstruction
Huiwon Jang, Sihyun Yu, Jinwoo Shin, Pieter Abbeel, Younggyo Seo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22853-22863
Efficient tokenization of videos remains a challenge in training vision models t hat can process long videos. One promising direction is to develop a tokenizer t hat can encode long video clips, as it would enable the tokenizer to leverage th e temporal coherence of videos better for tokenization. However, training existi ng tokenizers on long videos often incurs a huge training cost as they are train ed to reconstruct all the frames at once. In this paper, we introduce CoordTok, a video tokenizer that learns a mapping from coordinate-based representations to the corresponding patches of input videos, inspired by recent advances in 3D ge nerative models. In particular, CoordTok encodes a video into factorized triplan e representations and reconstructs patches that correspond to randomly sampled (x,y,t) coordinates. This allows for training large tokenizer models directly on long videos without requiring excessive training resources. Our experiments show that CoordTok can drastically reduce the number of tokens for encoding long vid eo clips. For instance, CoordTok can encode a 128-frame video with 128x128 resol ution into 1280 tokens, while baselines need 6144 or 8192 tokens to achieve simi lar reconstruction quality. We further show that this efficient video tokenizati on enables memory-efficient training of a diffusion transformer that can generat e 128 frames at once.
****

Derivative-Free Diffusion Manifold-Constrained Gradient for Unified XAI
Won Jun Kim, Hyungjin Chung, Jaemin Kim, Sangmin Lee, Byeongsu Sim, Jong Chul Ye ; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23795-23805
Gradient-based methods are a prototypical family of "explainability for AI" (XAI) techniques, especially for image-based models. However, they (1) require white -box access to models, (2) are vulnerable to adversarial attacks, and (3) produc e attributions that lie off the image manifold, leading to explanations that are not amenable to human perception. To overcome these challenges, we introduce De rivative-Free Diffusion Manifold-Contrained Gradients (FreeMCG): by leveraging e nsemble Kalman filters and diffusion models, we derive a derivative-free approxi mation of the model's gradient projected onto the data manifold, requiring acces s only to the model's outputs (i.e., black-box setting). We demonstrate the effe ctiveness of FreeMCG by applying it to both counterfactual generation and featur e attribution, which have traditionally been treated as different tasks requirin g distinct methods. Through comprehensive evaluation on both counterfactual expl anation and feature attribution we show that our method yields state-of-the-art results for both tasks while preserving the essential properties expected of XAI tools. Code: https://github.com/one-june/FreeMCG.
****

ZoomLDM: Latent Diffusion Model for Multi-scale Image Generation
Srikar Yellapragada, Alexandros Graikos, Kostas Triaridis, Prateek Prasanna, Raj arsi Gupta, Joel Saltz, Dimitris Samaras; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23453-23463
Diffusion models have revolutionized image generation, yet several challenges re strict their application to large-image domains, such as digital pathology and s

atellite imagery. Given that it is infeasible to directly train a model on 'whol e' images from domains with potential gigapixel sizes, diffusion-based generativ e methods have focused on synthesizing small, fixed-size patches extracted from these images. However, generating small patches has limited applicability since patch-based models fail to capture the global structures and wider context of la rge images, which can be crucial for synthesizing (semantically) accurate sample s. To overcome this limitation, we present ZoomLDM, a diffusion model tailored f or generating images across multiple scales. Central to our approach is a novel magnification-aware conditioning mechanism that utilizes self-supervised learnin g (SSL) embeddings and allows the diffusion model to synthesize images at differ ent 'zoom' levels, i.e., fixed-size patches extracted from large images at varyi ng scales. ZoomLDM synthesizes coherent histopathology images that remain contex tually accurate and detailed at different zoom levels, achieving state-of-the-ar t image generation quality across all scales and excelling in the data-scarce se tting of generating thumbnails of entire large images. The multi-scale nature of ZoomLDM unlocks additional capabilities in large image generation, enabling com putationally tractable and globally coherent image synthesis up to 4096x4096 pix els and 4x super-resolution. Additionally, multi-scale features extracted from Z oomLDM are highly effective in multiple instance learning experiments.

****

Do Computer Vision Foundation Models Learn the Low-level Characteristics of the Human Visual System?

Yancheng Cai, Fei Yin, Dounia Hammou, Rafal Mantiuk; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20039-20048

Computer vision foundation models, such as DINO or OpenCLIP, are trained in a se lf-supervised manner on large image datasets. Analogously, substantial evidence suggests that the human visual system (HVS) is influenced by the statistical dis tribution of colors and patterns in the natural world, characteristics also pres ent in the training data of foundation models. The question we address in this p aper is whether foundation models trained on natural images mimic some of the lo w-level characteristics of the human visual system, such as contrast detection, contrast masking, and contrast constancy. Specifically, we designed a protocol c omprising nine test types to evaluate the image encoders of 45 foundation and ge nerative models. Our results indicate that some foundation models (e.g., DINO, D INOv2, and OpenCLIP), share some of the characteristics of human vision, but oth er models show little resemblance. Foundation models tend to show smaller sensit ivity to low contrast and rather irregular responses to contrast across frequenc ies. The foundation models show the best agreement with human data in terms of c ontrast masking. Our findings suggest that human vision and computer vision may take both similar and different paths when learning to interpret images of the r eal world. Overall, while differences remain, foundation models trained on visio n tasks start to align with low-level human vision, with DINOv2 showing the clos est resemblance. Our code is available on https://github.com/caiyancheng/VFM_HVS _CVPR2025.

****

GaussianUDF: Inferring Unsigned Distance Functions through 3D Gaussian Splatting

Shujuan Li, Yu-Shen Liu, Zhizhong Han; Proceedings of the Computer Vision and Pa ttern Recognition Conference (CVPR), 2025, pp. 27113-27123

Reconstructing open surfaces from multi-view images is vital in digitalizing com plex objects in daily life. A widely used strategy is to learn unsigned distance functions (UDFs) by checking if their appearance conforms to the image observat ions through neural rendering. However, it is still hard to learn continuous and implicit UDF representations through 3D Gaussians splatting (3DGS) due to the d iscrete and explicit scene representation, i.e., 3D Gaussians. To resolve this i ssue, we propose a novel approach to bridge the gap between 3D Gaussians and UDF s. Our key idea is to overfit thin and flat 2D Gaussian planes on surfaces, and then, leverage the self-supervision and gradient-based inference to supervise un signed distances in both near and far area to surfaces. To this end, we introduc e novel constraints and strategies to constrain the learning of 2D Gaussians to pursue more stable optimization and more reliable self-supervision, addressing t

he challenges brought by complicated gradient field on or near the zero level se
t of UDFs. We report numerical and visual comparisons with the state-of-the-art
on widely used benchmarks and real data to show our advantages in terms of accur
acy, efficiency, completeness, and sharpness of reconstructed open surfaces with
 boundaries.
****

Towards RAW Object Detection in Diverse Conditions
Zhong-Yu Li, Xin Jin, Bo-Yuan Sun, Chun-Le Guo, Ming-Ming Cheng; Proceedings of
the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8859-88
68
Existing object detection methods often consider sRGB input, which was compresse
d from RAW data using ISP originally designed for visualization. However, such c
ompression might lose crucial information for detection, especially under comple
x light and weather conditions. We introduce the AODRaw dataset, which offers 7,
785 high-resolution real RAW images with 135,601 annotated instances spanning 62
 categories, capturing a broad range of indoor and outdoor scenes under 9 distin
ct light and weather conditions. Based on AODRaw that supports RAW and sRGB obje
ct detection, we provide a comprehensive benchmark for evaluating current detect
ion methods. We find that sRGB pre-training constrains the potential of RAW obje
ct detection due to the domain gap between sRGB and RAW, prompting us to directl
y pre-train on the RAW domain. However, it is harder for RAW pre-training to le
arn rich representations than sRGB pre-training due to the camera noise. To assi
st RAW pre-training, we distill the knowledge from an off-the-shelf model pre-tr
ained on the sRGB domain. As a result, we achieve substantial improvements under
 diverse and adverse conditions without relying on extra pre-processing modules.
 The code and dataset are available at https://github.com/lzyhha/AODRaw.
****

FLAME: Frozen Large Language Models Enable Data-Efficient Language-Image Pre-tra
ining
Anjia Cao, Xing Wei, Zhiheng Ma; Proceedings of the Computer Vision and Pattern
Recognition Conference (CVPR), 2025, pp. 4080-4090
Language-image pre-training faces significant challenges due to limited data in
specific formats and the constrained capacities of text encoders. While prevaili
ng methods attempt to address these issues through data augmentation and archite
cture modifications, they continue to struggle with processing long-form text in
puts, and the inherent limitations of traditional CLIP text encoders lead to sub
optimal downstream generalization. In this paper, we propose FLAME (Frozen Large
 lAnguage Models Enable data-efficient language-image pre-training) that leverag
es frozen large language models as text encoders, naturally processing long text
 inputs and demonstrating impressive multilingual generalization. FLAME comprise
s two key components: 1) a multifaceted prompt distillation technique for extrac
ting diverse semantic representations from long captions, which better aligns wi
th the multifaceted nature of images, and 2) a facet-decoupled attention mechani
sm, complemented by an offline embedding strategy, to ensure efficient computati
on. Extensive empirical evaluations demonstrate FLAME's superior performance. Wh
en trained on CC3M, FLAME surpasses the previous state-of-the-art by 4.9% in Ima
geNet top-1 accuracy. On YFCC15M, FLAME surpasses the WIT-400M-trained CLIP by 4
4.4% in average image-to-text recall@1 across 36 languages, and by 34.6% in text
-to-image recall@1 for long-context retrieval on Urban-1k. Code is available at
https://github.com/MIV-XJTU/FLAME.
****

CrossSDF: 3D Reconstruction of Thin Structures From Cross-Sections
Thomas Walker, Salvatore Esposito, Daniel Rebain, Amir Vaxman, Arno Onken, Chang
jian Li, Oisin Mac Aodha; Proceedings of the Computer Vision and Pattern Recogni
tion Conference (CVPR), 2025, pp. 30928-30937
Reconstructing complex structures from planar cross-sections is a challenging pr
oblem, with wide-reaching applications in medical imaging, manufacturing, and to
pography. Out-of-the-box point cloud reconstruction methods can often fail due t
o the data sparsity between slicing planes, while current bespoke methods strugg
le to reconstruct thin geometric structures and preserve topological continuity.

This is important for medical applications where thin vessel structures are present in CT and MRI scans. This paper introduces CrossSDF, a novel approach for extracting a 3D signed distance field from 2D signed distances generated from planar contours. Our approach makes the training of neural SDFs contour-aware by using losses designed for the case where geometry is known within 2D slices. Our results demonstrate a significant improvement over existing methods, effectively reconstructing thin structures and producing accurate 3D models without the interpolation artifacts or over-smoothing of prior approaches.
****

DV-Matcher: Deformation-based Non-rigid Point Cloud Matching Guided by Pre-trained Visual Features
Zhangquan Chen, Puhua Jiang, Ruqi Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27264-27274
In this paper, we present DV-Matcher, a novel learning-based framework for estimating dense correspondences between non-rigidly deformable point clouds. Learning directly from unstructured point clouds without meshing or manual labelling, our framework delivers high-quality dense correspondences, which is of significant practical utility in point cloud processing. Our key contributions are two-fold: First, we propose a scheme to inject prior knowledge from pre-trained vision models into geometric feature learning, which effectively complements the local nature of geometric features with global and semantic information; Second, we propose a novel deformation-based module to promote the extrinsic alignment induced by the learned correspondences, which effectively enhances the feature learning. Experimental results show that our method achieves state-of-the-art results in matching non-rigid point clouds in both near-isometric and heterogeneous shape collection as well as more realistic partial and noisy data. Our code is available at https://github.com/rqhuang88/DV-Matcher.
****

Reasoning Mamba: Hypergraph-Guided Region Relation Calculating for Weakly Supervised Affordance Grounding
Yuxuan Wang, Aming Wu, Muli Yang, Yukuan Min, Yihang Zhu, Cheng Deng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27618-27627
This paper pays attention to Weakly Supervised Affordance Grounding (WSAG) task that aims to train model to identify affordance regions using human-object interaction images and egocentric images without the need for costly pixel-level annotations. Most existing methods usually consider the affordance regions to be isolated and directly employ class activation maps to conduct localization, ignoring the relationships with other object components and weakening the performance. For example, a cup's handle is combined with its body to achieve the pouring ability. Obviously, capturing the region relationships is beneficial for improving the localization accuracy of affordance regions. To this end, we first explore exploiting hypergraph to discover these relations and propose a Reasoning Mamba (R-Mamba) framework. We first extract feature embeddings from exo-centric and egocentric images to construct the hypergraphs consisting of multiple vertices and hyperedges, which capture the in-context local region relationships between different visual components. Subsequently, we design a Hypergraph-guided State Space (HSS) block to reorganize these local relationships from the global perspective. By this mechanism, the model could leverage the captured relationships to improve the localization accuracy of affordance regions. Extensive experiments and visualization analyses demonstrate the superiority of our method.
****

Adapter Merging with Centroid Prototype Mapping for Scalable Class-Incremental Learning
Takuma Fukuda, Hiroshi Kera, Kazuhiko Kawamoto; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4884-4893
We propose Adapter Merging with Centroid Prototype Mapping (ACMap), an exemplar-free framework for class-incremental learning (CIL) that addresses both catastrophic forgetting and scalability. While existing methods trade-off between inference time and accuracy, ACMap consolidates task-specific adapters into a single a

dapter, ensuring constant inference time across tasks without compromising accuracy.The framework employs adapter merging to build a shared subspace that aligns task representations and mitigates forgetting, while centroid prototype mapping maintains high accuracy through consistent adaptation in the shared subspace. To further improve scalability, an early stopping strategy limits adapter merging as tasks increase. Extensive experiments on five benchmark datasets demonstrate that ACMap matches state-of-the-art accuracy while maintaining inference speed comparable to the fastest existing methods.
****

OpenSDI: Spotting Diffusion-Generated Images in the Open World
Yabin Wang, Zhiwu Huang, Xiaopeng Hong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4291-4301
This paper identifies OpenSDI, a challenge for spotting diffusion-generated images in open-world settings. In response to this challenge, we define a new benchmark, the OpenSDI dataset (OpenSDID), which stands out from existing datasets due to its diverse use of large vision-language models that simulate open-world diffusion-based manipulations. Another outstanding feature of OpenSDID is its inclusion of both detection and localization tasks for images manipulated globally and locally by diffusion models. To address the OpenSDI challenge, we propose a Synergizing Pretrained Models (SPM) scheme to build up a mixture of foundation models. This approach exploits a collaboration mechanism with multiple pretrained foundation models to enhance generalization in the OpenSDI context, moving beyond traditional training by synergizing multiple pretrained models through prompting and attending strategies. Building on this scheme, we introduce MaskCLIP, an SPM-based model that aligns Contrastive Language-Image Pre-Training (CLIP) with Masked Autoencoder (MAE). Extensive evaluations on OpenSDID show that MaskCLIP significantly outperforms current state-of-the-art methods for the OpenSDI challenge, achieving remarkable relative improvements of 14.23% in IoU (14.11% in F1) and 2.05% in accuracy (2.38% in F1) compared to the second-best model in localization and detection tasks, respectively. Our dataset and code are available at https://github.com/iamwangyabin/OpenSDI.
****

Adaptive Part Learning for Fine-Grained Generalized Category Discovery: A Plug-and-Play Enhancement
Qiyuan Dai, Hanzhuo Huang, Yu Wu, Sibei Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25444-25453
Generalized Category Discovery (GCD) aims to recognize unlabeled images from known and novel classes by distinguishing novel classes from known ones, while also transferring knowledge from another set of labeled images with known classes. Existing GCD methods rely on self-supervised vision transformers such as DINO for representation learning. However, focusing solely on the global representation of the DINO CLS token introduces an inherent trade-off between discriminability and generalization. In this paper, we introduce an adaptive part discovery and learning method, called APL, which generates consistent object parts and their correspondences across different similar images using a set of shared learnable part queries and DINO part priors, without requiring any additional annotations. More importantly, we propose a novel all-min contrastive loss to learn discriminative yet generalizable part representation, which adaptively highlights discriminative object parts to distinguish similar categories for enhanced discriminability while simultaneously sharing other parts to facilitate knowledge transfer for improved generalization. Our APL can easily be incorporated into different GCD frameworks by replacing their CLS token feature with our part representations, showing significant enhancements on fine-grained datasets.
****

Online Task-Free Continual Learning via Dynamic Expansionable Memory Distribution
Fei Ye, Adrian G. Bors; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20512-20522
Recent continuous learning (CL) research primarily addresses catastrophic forgetting within a straightforward learning framework where class and task informatio

n are predefined. However, in the Task-Free Continual Learning (TFCL), representing a more realistic and challenging CL scenarios, such information is typically absent. In this paper, we address the online TFCL by introducing an innovative memory management approach, by incorporating a dynamic memory system for storing selected data representatives from evolving distributions while a dynamically expandable memory system enables the retention of essential long-term knowledge. The proposed dynamic expandable memory system manages a series of memory distributions, each designed to represent the information from a distinct data category. A new memory expansion mechanism that assesses the proximity between incoming samples and existing memory distributions is proposed for evaluating when to add new memory distributions into the system. Additionally, a novel memory distribution augmentation technique is proposed for selectively gathering suitable samples for each memory distribution, enhancing the statistical robustness over time. To prevent memory saturation before the training phase, we introduce a memory distribution reduction strategy that automatically eliminates overlapping memory distributions, ensuring adequate capacity for accommodating new information in subsequent learning episodes. We conduct a series of experiments demonstrating that our proposed approach attains state-of-the-art performance in both supervised and unsupervised learning contexts. The source code is available at https://github.com/dtuzi123/DEMD.

****

Lux Post Facto: Learning Portrait Performance Relighting with Conditional Video Diffusion and a Hybrid Dataset

Yiqun Mei, Mingming He, Li Ma, Julien Philip, Wenqi Xian, David M George, Xueming Yu, Gabriel Dedic, Ahmet Levent Ta■el, Ning Yu, Vishal M. Patel, Paul Debevec; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5510-5522

Video portrait relighting remains challenging because the results need to be both photorealistic and temporally stable.This typically requires a strong model design that can capture complex facial reflections as well as intensive training on a high-quality paired video dataset, such as dynamic one-light-at-a-time (OLAT). In this work, we introduce Lux Post Facto, a novel portrait video relighting method that produces both photorealistic and temporally consistent lighting effects. From the model side, we design a new conditional video diffusion model built upon state-of-the-art pre-trained video diffusion model, alongside a new lighting injection mechanism to enable precise control. This way we leverage strong spatial and temporal generative capability to generate plausible solutions to the ill-posed relighting problem. Our technique uses a hybrid dataset consisting of static expression OLAT data and in-the-wild portrait performance videos to jointly learn relighting and temporal modeling. This avoids the need to acquire paired video data in different lighting conditions. Our extensive experiments show that our model produces state-of-the-art results both in terms of photorealism and temporal consistency.

****

DiG: Scalable and Efficient Diffusion Models with Gated Linear Attention

Lianghui Zhu, Zilong Huang, Bencheng Liao, Jun Hao Liew, Hanshu Yan, Jiashi Feng, Xinggang Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7664-7674

Diffusion models with large-scale pre-training have achieved significant success in the field of visual content generation, particularly exemplified by Diffusion Transformers (DiT). However, DiT models have faced challenges with quadratic complexity efficiency, especially when handling long sequences. In this paper, we aim to incorporate the sub-quadratic modeling capability of Gated Linear Attention (GLA) into the 2D diffusion backbone. Specifically, we introduce Diffusion Gated Linear Attention Transformers (DiG), a simple, adoptable solution with minimal parameter overhead. We offer two variants, i.e., a plain and U-shape architecture, showing superior efficiency and competitive effectiveness. In addition to superior performance to DiT and other sub-quadratic-time diffusion models at 256x256 resolution, DiG demonstrates greater efficiency than these methods starting from a 512 resolution. Specifically, DiG-S/2 is 2.5x faster and saves 75.7% GP

U memory compared to DiT-S/2 at a 1792 resolution. Additionally, DiG-XL/2 is 4.2 x faster than the Mamba-based model at a 1024 resolution and 1.8x faster than Di T with FlashAttention-2 at a 2048 resolution. The code is released at https://gi thub.com/hustvl/DiG.
****
Monocular and Generalizable Gaussian Talking Head Animation
Shengjie Gong, Haojie Li, Jiapeng Tang, Dongming Hu, Shuangping Huang, Hao Chen, Tianshui Chen, Zhuoman Liu; Proceedings of the Computer Vision and Pattern Reco gnition Conference (CVPR), 2025, pp. 5523-5534
In this work, we introduce Monocular and Generalizable Gaussian Talking Head Ani mation (MGGTalk), which requires monocular datasets and generalizes to unseen id entities without personalized re-training. Compared with previous 3D Gaussian Sp latting (3DGS) methods that requires elusive multi-view datasets or tedious pers onalized learning/inference, MGGtalk enables more practical and broader applicat ions. However, in the absence of multi-view and personalized training data, the incompleteness of geometric and appearance information poses a significant chall enge. To address these challenges, MGGTalk explores depth information to enhance geometric and facial symmetry characteristics to supplement both geometric and appearance features. Initially, based on the pixel-wise geometric information ob tained from depth estimation, we incorporate symmetry operations and point cloud filtering techniques to ensure a complete and precise position parameter for 3D GS. Subsequently, we adopt a two-stage strategy with symmetric priors for predic ting the remaining 3DGS parameters. We begin by predicting Gaussian parameters f or the visible facial regions of the source image. These parameters are subseque ntly utilized to improve the prediction of Gaussian parameters for the non-visib le regions. Extensive experiments demonstrate that MGGTalk surpasses previous st ate-of-the-art methods, achieving superior performance across various metrics.
****
Rethinking Token Reduction with Parameter-Efficient Fine-Tuning in ViT for Pixel -Level Tasks
Cheng Lei, Ao Li, Hu Yao, Ce Zhu, Le Zhang; Proceedings of the Computer Vision a nd Pattern Recognition Conference (CVPR), 2025, pp. 14954-14964
Parameter-efficient fine-tuning (PEFT) adapts pre-trained models to new tasks by updating only a small subset of parameters, achieving efficiency but still faci ng significant inference costs driven by input token length. This challenge is e ven more pronounced in pixel-level tasks, which require longer input sequences c ompared to image-level tasks. Although token reduction (TR) techniques can help reduce computational demands, they often lead to homogeneous attention patterns that compromise performance in pixel-level scenarios. This study underscores the importance of maintaining attention diversity for these tasks and proposes to e nhance attention diversity while ensuring the completeness of token sequences. O ur approach effectively reduces the number of tokens processed within transforme r blocks, improving computational efficiency without sacrificing performance on several pixel-level tasks. We also demonstrate the superior generalization capab ility of our proposed method compared to challenging baseline models. The source code will be made available at https://github.com/AVC2-UESTC/DAR-TR-PEFT.
****
SVDC: Consistent Direct Time-of-Flight Video Depth Completion with Frequency Sel ective Fusion
Xuan Zhu, Jijun Xiang, Xianqi Wang, Longliang Liu, Yu Wang, Hong Zhang, Fei Guo, Xin Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16619-16628
Lightweight direct Time-of-Flight (dToF) sensors are ideal for 3D sensing on mob ile devices. However, due to the manufacturing constraints of compact devices an d the inherent physical principles of imaging, dToF depth maps are sparse and no isy. In this paper, we propose a novel video depth completion method, called SVD C, by fusing the sparse dToF data with the corresponding RGB guidance. Our metho d employs a multi-frame fusion scheme to mitigate the spatial ambiguity resultin g from the sparse dToF imaging. Misalignment between consecutive frames during m ulti-frame fusion could cause blending between object edges and the background,

which results in a loss of detail. To address this, we introduce an adaptive frequency selective fusion (AFSF) module, which automatically selects convolution kernel sizes to fuse multi-frame features. Our AFSF utilizes a channel-spatial enhancement attention (CSEA) module to enhance features and generates an attention map as fusion weights. The AFSF ensures edge detail recovery while suppressing high-frequency noise in smooth regions. To further enhance temporal consistency, We propose a cross-window consistency loss to ensure consistent predictions across different windows, effectively reducing flickering. Our proposed SVDC achieves optimal accuracy and consistency on the TartanAir and Dynamic Replica datasets. Code is available at https://github.com/Lan1eve/SVDC.
****

Locally Orderless Images for Optimization in Differentiable Rendering
Ishit Mehta, Manmohan Chandraker, Ravi Ramamoorthi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5763-5772
Problems in differentiable rendering often involve optimizing scene parameters that cause motion in image space. The gradients for such parameters tend to be sparse, leading to poor convergence. While existing methods address this sparsity through proxy gradients such as topological derivatives or lagrangian derivatives, they make simplifying assumptions about rendering. Multi-resolution image pyramids offer an alternative approach but prove unreliable in practice. We introduce a method that uses locally orderless images --- where each pixel maps to a histogram of intensities that preserves local variations in appearance. Using an inverse rendering objective that minimizes histogram distance, our method extends support for sparsely defined image gradients and recovers optimal parameters. We validate our method on various inverse problems using both synthetic and real data.
****

Plug-and-Play Interpretable Responsible Text-to-Image Generation via Dual-Space Multi-facet Concept Control
Basim Azam, Naveed Akhtar; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2976-2985
Ethical issues around text-to-image (T2I) models demand a comprehensive control over the generative content. Existing techniques addressing these issues for responsible T2I models aim for the generated content to be fair and safe (non-violent/explicit). However, these methods remain bounded to handling the facets of responsibility concepts individually, while also lacking in interpretability. Moreover, they often require alteration to the original model, which compromises the model performance. In this work, we propose a unique technique to enable responsible T2I generation by simultaneously accounting for an extensive range of concepts for fair and safe content generation in a scalable manner. The key idea is to distill the target T2I pipeline with an external plug-and-play mechanism that learns an interpretable composite responsible space for the desired concepts, conditioned on the target T2I pipeline. We use knowledge distillation and concept whitening to enable this. At inference, the learned space is utilized to modulate the generative content. A typical T2I pipeline presents two plug-in points for our approach, namely; the text embedding space and the diffusion model latent space. We develop modules for both points and show the effectiveness of our approach with a range of strong results. Our code and models will be made public after paper acceptance.
****

Rethinking Training for De-biasing Text-to-Image Generation: Unlocking the Potential of Stable Diffusion
Eunji Kim, Siwon Kim, Minjun Park, Rahim Entezari, Sungroh Yoon; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13361-13370
Recent advancements in text-to-image models, such as Stable Diffusion, show significant demographic biases. Existing de-biasing techniques rely heavily on additional training, which imposes high computational costs and risks of compromising core image generation functionality. This hinders them from being widely adopted to real-world applications. In this paper, we explore Stable Diffusion's overl

ooked potential to reduce bias without requiring additional training. Through ou
r analysis, we uncover that initial noises associated with minority attributes f
orm "minority regions" rather than scattered. We view these "minority regions" a
s opportunities in SD to reduce bias. To unlock the potential, we propose a nove
l de-biasing method called 'weak guidance,' carefully designed to guide a random
 noise to the minority regions without compromising semantic integrity. Through
analysis and experiments on various versions of SD, we demonstrate that our prop
osed approach effectively reduces bias without additional training, achieving bo
th efficiency and preservation of core image generation functionality.
****

FLAIR: VLM with Fine-grained Language-informed Image Representations
Rui Xiao, Sanghwan Kim, Mariana-Iuliana Georgescu, Zeynep Akata, Stephan Alaniz;
 Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2
025, pp. 24884-24894
CLIP has shown impressive results in aligning images and text at scale. However,
 its ability to capture detailed visual features remains limited because CLIP ma
tches images and texts at a global level. To address this issue, we propose FLAI
R, Fine-grained Language-informed Image Representations, an approach that utiliz
es long and detailed image descriptions to learn localized image embeddings. By
sampling diverse sub-captions that describe fine-grained details about an image,
 we train our vision-language model to produce not only global embeddings but al
so text-specific image representations. Our model introduces text-conditioned at
tention pooling on top of local image tokens to produce fine-grained image repre
sentations that excel at retrieving detailed image content. We achieve state-of-
the-art performance on both, existing multimodal retrieval benchmarks, as well a
s, our newly introduced fine-grained retrieval task which evaluates vision-langu
age models' ability to retrieve partial image content. Furthermore, our experime
nts demonstrate the effectiveness of FLAIR trained on 30M image-text pairs in ca
pturing fine-grained visual information, including zero-shot semantic segmentati
on, outperforming models trained on billions of pairs. Code is available at: htt
ps://github.com/ExplainableML/flair.
****

GG-SSMs: Graph-Generating State Space Models
Nikola Zubic, Davide Scaramuzza; Proceedings of the Computer Vision and Pattern
Recognition Conference (CVPR), 2025, pp. 28863-28873
State Space Models (SSMs) are powerful tools for modeling sequential data in com
puter vision and time series analysis domains. However, traditional SSMs are lim
ited by fixed, one-dimensional sequential processing, which restricts their abil
ity to model non-local interactions in high-dimensional data. While methods like
 Mamba and VMamba introduce selective and flexible scanning strategies, they rel
y on predetermined paths, which fails to efficiently capture complex dependencie
s. We introduce Graph-Generating State Space Models (GG-SSMs), a novel framework
 that overcomes these limitations by dynamically constructing graphs based on fe
ature relationships. Using Chazelle's Minimum Spanning Tree algorithm, GG-SSMs a
dapt to the inherent data structure, enabling robust feature propagation across
dynamically generated graphs and efficiently modeling complex dependencies. We v
alidate GG-SSMs on 11 diverse datasets, including event-based eye-tracking, Imag
eNet classification, optical flow estimation, and six time series datasets. GG-S
SMs achieve state-of-the-art performance across all tasks, surpassing existing m
ethods by significant margins. Specifically, GG-SSM attains a top-1 accuracy of
84.9% on ImageNet, outperforming prior SSMs by 1%, reducing the KITTI-15 error r
ate to 2.77%, and improving eye-tracking detection rates by up to 0.33% with few
er parameters. These results demonstrate that dynamic scanning based on feature
relationships significantly improves SSMs' representational power and efficiency
, offering a versatile tool for various applications in computer vision and beyo
nd.
****

Instant3dit: Multiview Inpainting for Fast Editing of 3D Objects
Amir Barda, Matheus Gadelha, Vladimir G. Kim, Noam Aigerman, Amit H. Bermano, Th
ibault Groueix; Proceedings of the Computer Vision and Pattern Recognition Confe

rence (CVPR), 2025, pp. 16273-16282
We propose a generative technique to edit 3D shapes, represented as meshes, NeRFs, or Gaussian Splats, in ~3 seconds, without the need for running an SDS type of optimization.Our key insight is to cast 3D editing as a multiview image inpainting problem, as this representation is generic and can be mapped back to any 3D representation using the bank of available Large Reconstruction Models. We explore different fine-tuning strategies to obtain both multiview generation and inpainting capabilities within the same diffusion model. In particular, the design of the inpainting mask is an important factor of training an inpainting model, and we propose several masking strategies to mimic the types of edits a user would perform on a 3D shape. Our approach takes 3D generative editing from hours to seconds and produces higher-quality results compared to previous works.
****

STDD: Spatio-Temporal Dual Diffusion for Video Generation
Shuaizhen Yao, Xiaoya Zhang, Xin Liu, Mengyi Liu, Zhen Cui; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12575-12584
Diffusion probabilistic model is becoming the cornerstone of data generation, especially generating high-quality images. As an extension, video diffusion generation is in urgent need of a principled temporal-sequence diffusion way, while the spatial-domain diffusion dominates most video diffusion methods. In this work, we propose an explicit Spatio-Temporal Dual Diffusion (STDD) method by principledly extending the standard diffusion model to a spatio-temporal diffusion model for joint spatial and temporal noise propagation/reduction. Mathematically, an analysable dual diffusion process is derived to accumulate noises/information in temporal sequence as well as spatial domain. Correspondingly, we theoretically derive a spatio-temporal probabilistic reverse diffusion process and propose an accelerated sampling way to reduce the inference cost. In principle, the spatio-temporal dual diffusion enables the information of previous frames to be transferred to the current frame, which thus could be beneficial for video consistency. Extensive experiments demonstrate that our proposed STDD is more competitive over the state-of-the-art methods in the task of video generation/prediction as well as text-to-video generation.
****

Implicit Correspondence Learning for Image-to-Point Cloud Registration
Xinjun Li, Wenfei Yang, Jiacheng Deng, Zhixin Cheng, Xu Zhou, Tianzhu Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16922-16931
Image-to-point cloud registration aims to estimate the camera pose of a given image within a 3D scene point cloud. In this area, matching-based methods have achieved leading performance by first detecting the overlapping region, then matching point and pixel features learned by neural networks and finally using the PnP-RANSAC algorithm to estimate camera pose. However, achieving accurate image-to-point cloud registration remains challenging because the overlapping region detection is unreliable merely relying on point-wise classification, direct alignment of cross-modal data is difficult and indirect optimization objective leads to unstable registration results. To address these challenges, we propose a novel implicit correspondence learning method, including a Geometric Prior-guided overlapping region Detection Module (GPDM), an Implicit Correspondence Learning Module (ICLM), and a Pose Regression Module (PRM). The proposed method enjoys several merits. First, the proposed GPDM can precisely detect the overlapping region. Second, the ICLM can generate robust cross-modality correspondences. Third, the PRM can enable end-to-end optimization. Extensive experimental results on KITTI and nuScenes datasets demonstrate that the proposed model sets a new state-of-the-art performance in registration accuracy.
****

Continuous Adverse Weather Removal via Degradation-Aware Distillation
Xin Lu, Jie Xiao, Yurui Zhu, Xueyang Fu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28113-28123
All-in-one models for adverse weather removal aim to process various degraded images using a single set of parameters, making them ideal for real-world scenario

s. However, they encounter two main challenges: catastrophic forgetting and limi ted degradation awareness. The former causes the model to lose knowledge of prev iously learned scenarios, reducing its overall effectiveness. While the later ha mpers the model's ability to accurately identify and respond to specific types o f degradation, limiting its performance across diverse adverse weather condition s. To address these issues, we introduce the Incremental Learning Adverse Weathe r Removal (ILAWR) framework, which uses a novel degradation-aware distillation s trategy for continuous weather removal. Specifically, we first design a degradat ion-aware module that utilizes Fourier priors to capture a broad range of degrad ation features, effectively mitigating catastrophic forgetting in low-level visu al tasks. Then, we implement multilateral distillation, which combines knowledge from multiple teacher models using an importance-guided aggregation approach. T his enables the model to balance adaptation to new degradation types with the pr eservation of background details. Extensive experiments on both synthetic and re al-world datasets confirm that ILAWR outperforms existing models across multiple benchmarks, proving its effectiveness in continuous adverse weather removal.
****

Fine-Grained Erasure in Text-to-Image Diffusion-based Foundation Models
Kartik Thakral, Tamar Glaser, Tal Hassner, Mayank Vatsa, Richa Singh; Proceeding s of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 91 21-9130
Existing unlearning algorithms in text-to-image generative models often fail to preserve the knowledge of semantically related concepts when removing specific t arget concepts--a challenge known as adjacency. To address this, we propose FADE (Fine-grained Attenuation for Diffusion Erasure), introducing adjacency-aware u nlearning in diffusion models. FADE comprises two components: (1) the Concept Ne ighborhood, which identifies an adjacency set of related concepts, and (2) Mesh Modules, employing a structured combination of Expungement, Adjacency, and Guida nce loss components. These enable precise erasure of target concepts while prese rving fidelity across related and unrelated concepts. Evaluated on datasets like Stanford Dogs, Oxford Flowers, CUB, I2P, Imagenette, and ImageNet-1k, FADE effe ctively removes target concepts with minimal impact on correlated concepts, achi eving at least a 12% improvement in retention performance over state-of-the-art methods. Our code and models are available on the project page: iab-rubric/unlea rning/FG-Un.
****

ILIAS: Instance-Level Image retrieval At Scale
Giorgos Kordopatis-Zilos, Vladan Stojni■, Anna Manko, Pavel Suma, Nikolaos-Anton ios Ypsilantis, Nikos Efthymiadis, Zakaria Laskar, Jiri Matas, Ondrej Chum, Gior gos Tolias; Proceedings of the Computer Vision and Pattern Recognition Conferenc e (CVPR), 2025, pp. 14777-14787
This work introduces ILIAS, a new test dataset for Instance-Level Image retrieva l At Scale. It is designed to evaluate the ability of current and future foundat ion models and retrieval techniques to recognize particular objects. The key ben efits over existing datasets include large scale, domain diversity, accurate gro und truth, and a performance that is far from saturated. ILIAS includes query an d positive images for 1,000 object instances, manually collected to capture chal lenging conditions and diverse domains. Large-scale retrieval is conducted again st 100 million distractor images from YFCC100M. To avoid false negatives without extra annotation effort, we include only query objects confirmed to have emerge d after 2014, i.e. the compilation date of YFCC100M. An extensive benchmarking i s performed with the following observations: i) models fine-tuned on specific do mains, such as landmarks or products, excel in that domain but fail on ILIAS ii) learning a linear adaptation layer using multi-domain class supervision results in performance improvements, especially for vision-language models iii) local d escriptors in retrieval re-ranking are still a key ingredient, especially in the presence of severe background clutter iv) the text-to-image performance of the vision-language foundation models is surprisingly close to the corresponding ima ge-to-image case. website: https://vrg.fel.cvut.cz/ilias/
****

Exploiting Temporal State Space Sharing for Video Semantic Segmentation
Syed Ariff Syed Hesham, Yun Liu, Guolei Sun, Henghui Ding, Jing Yang, Ender Konukoglu, Xue Geng, Xudong Jiang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24211-24221
Video semantic segmentation (VSS) plays a vital role in understanding the temporal evolution of scenes. Traditional methods often segment videos frame-by-frame or in a short temporal window, leading to limited temporal context, redundant computations, and heavy memory requirements. To this end, we introduce a Temporal Video State Space Sharing (TV3S) architecture to leverage Mamba state space models for temporal feature sharing. Our model features a selective gating mechanism that efficiently propagates relevant information across video frames, eliminating the need for a memory-heavy feature pool. By processing spatial patches independently and incorporating shifted operation, TV3S supports highly parallel computation in both training and inference stages, which reduces the delay in sequential state space processing and improves the scalability for long video sequences. Moreover, TV3S incorporates information from prior frames during inference, achieving long-range temporal coherence and superior adaptability to extended sequences. Evaluations on the VSPW and Cityscapes datasets reveal that our approach outperforms current state-of-the-art methods, establishing a new standard for VSS with consistent results across long video sequences. By achieving a good balance between accuracy and efficiency, TV3S shows a significant advancement in spatiotemporal modeling, paving the way for efficient video analysis. The code is publicly available at https://github.com/Ashesham/TV3S.git.
****

DeRS: Towards Extremely Efficient Upcycled Mixture-of-Experts Models
Yongqi Huang, Peng Ye, Chenyu Huang, Jianjian Cao, Lin Zhang, Baopu Li, Gang Yu, Tao Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10056-10066
Upcycled Mixture-of-Experts (MoE) models have shown great potential in various tasks by converting the original Feed-Forward Network (FFN) layers in pre-trained dense models into MoE layers. However, these models still suffer from significant parameter inefficiency due to the introduction of multiple experts. In this work, we propose a novel DeRS (Decompose, Replace, and Synthesis) paradigm to overcome this shortcoming, which is motivated by our observations about the unique redundancy mechanisms of upcycled MoE experts. Specifically, DeRS decomposes the experts into one expert-shared base weight and multiple expert-specific delta weights, and subsequently represents these delta weights in lightweight forms. Our proposed DeRS paradigm can be applied to enhance parameter efficiency in two different scenarios, including: 1) DeRS Compression for inference stage, using sparsification or quantization to compress vanilla upcycled MoE models; and 2) DeRS Upcycling for training stage, employing lightweight sparse or low-rank matrixes to efficiently upcycle dense models into MoE models. Extensive experiments across three different tasks show that the proposed methods can achieve extreme parameter efficiency while maintaining the performance for both training and compression of upcycled MoE models.
****

GeoDepth: From Point-to-Depth to Plane-to-Depth Modeling for Self-Supervised Monocular Depth Estimation
Haifeng Wu, Shuhang Gu, Lixin Duan, Wen Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11525-11535
Self-supervised monocular depth estimation has long been treated as a point-wise prediction problem, where the depth of each pixel is usually estimated independently. However, artifacts are often observed in the estimated depth map, e.g., depth values for points located in the same region may jump dramatically. To address this issue, we propose a novel self-supervised monocular depth estimation framework called GeoDepth, where we explore the intrinsic geometric representation in 3D scenes for producing accurate and continuous depth maps. In particular, we model the complex 3D scene as a collection of planes with varying sizes, where each plane is characterized by a unique set of parameters, namely planar normal (indicating plane orientation) and planar offset (defining the perpendicular di

stance from the camera center to the plane). Under this modeling, points in the same plane are enforced to share a unique representation and their depth variations related only to pixel coordinates, thus this geometric relationship can be exploited to regularize the depth variations of these points. To this end, we design a structured plane generation module that introduces spatio-temporal geometric cues and the plane uniqueness principle to recover the correct scene plane representation. In addition, we develop a depth discontinuity module to identify depth discontinuity regions and subsequently optimize them. Our experiments on the KITTI and NYUv2 datasets demonstrate that GeoDepth achieves state-of-the-art performance, with additional tests on Make3D and ScanNet validating its generalization capabilities.

****

SSHNet: Unsupervised Cross-modal Homography Estimation via Problem Reformulation and Split Optimization

Junchen Yu, Si-Yuan Cao, Runmin Zhang, Chenghao Zhang, Zhu Yu, Shujie Chen, Bailin Yang, Hui-Liang Shen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16685-16694

We propose a novel unsupervised cross-modal homography estimation learning framework, named Split Supervised Homography estimation Network (SSHNet). SSHNet reformulates the unsupervised cross-modal homography estimation into two supervised sub-problems, each addressed by its specialized network: a homography estimation network and a modality transfer network. To realize stable training, we introduce an effective split optimization strategy to train each network separately within its respective sub-problem. We also formulate an extra homography feature space supervision to enhance feature consistency, further boosting the estimation accuracy. Moreover, we employ a simple yet effective distillation training technique to reduce model parameters and improve cross-domain generalization ability while maintaining comparable performance. The training stability of SSHNet enables its cooperation with various homography estimation architectures. Experiments reveal that the SSHNet using IHN as homography estimation network, namely SSHNet-IHN, outperforms previous unsupervised approaches by a significant margin. Even compared to supervised approaches MHN and LocalTrans, SSHNet-IHN achieves 47.4% and 85.8% mean average corner errors (MACEs) reduction on the challenging OPT-SAR dataset. Source code is available at https://github.com/Junchen-Yu/SSHNet.

****

High-fidelity 3D Object Generation from Single Image with RGBN-Volume Gaussian Reconstruction Model

Yiyang Shen, Kun Zhou, He Wang, Yin Yang, Tianjia Shao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21558-21569

Recently single-view 3D generation via Gaussian splatting has emerged and developed quickly. They learn 3D Gaussians from 2D RGB images generated from pre-trained multi-view diffusion (MVD) models, and have shown a promising avenue for 3D generation through a single image. Despite the current progress, these methods still suffer from the inconsistency jointly caused by the geometric ambiguity in the 2D images, and the lack of structure of 3D Gaussians, leading to distorted and blurry 3D object generation. In this paper, we propose to fix these issues by GS-RGBN, a new RGBN-volume Gaussian Reconstruction Model designed to generate high-fidelity 3D objects from single-view images. Our key insight is a structured 3D representation can simultaneously mitigate the afore-mentioned two issues. To this end, we propose a novel hybrid Voxel-Gaussian representation, where a 3D voxel representation contains explicit 3D geometric information, eliminating the geometric ambiguity from 2D images. It also structures Gaussians during learning so that the optimization tends to find better local optima. Our 3D voxel representation is obtained by a fusion module that aligns RGB features and surface normal features, both of which can be estimated from 2D images. Extensive experiments demonstrate the superiority of our methods over prior works in terms of high-quality reconstruction results, robust generalization, and good efficiency.

****

Steepest Descent Density Control for Compact 3D Gaussian Splatting

Peihao Wang, Yuehao Wang, Dilin Wang, Sreyas Mohan, Zhiwen Fan, Lemeng Wu, Ruisi

Cai, Yu-Ying Yeh, Zhangyang Wang, Qiang Liu, Rakesh Ranjan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26663-26672
3D Gaussian Splatting (3DGS) has emerged as a powerful technique for real-time, high-resolution novel view synthesis. By representing scenes as a mixture of Gaussian primitives, 3DGS leverages GPU rasterization pipelines for efficient rendering and reconstruction. To optimize scene coverage and capture fine details, 3DGS employs a densification algorithm to generate additional points. However, this process often leads to redundant point clouds, resulting in excessive memory usage, slower performance, and substantial storage demands - posing significant challenges for deployment on resource-constrained devices. To address this limitation, we propose a theoretical framework that demystifies and improves density control in 3DGS. Our analysis reveals that splitting is crucial for escaping saddle points. Through an optimization-theoretic approach, we establish the necessary conditions for densification, determine the minimal number of offspring Gaussians, identify the optimal parameter update direction, and provide an analytical solution for normalizing off-spring opacity. Building on these insights, we introduce SteepGS, incorporating steepest density control, a principled strategy that minimizes loss while maintaining a compact point cloud. SteepGS achieves a 50% reduction in Gaussian points without compromising rendering quality, significantly enhancing both efficiency and scalability.

****

Optimal Transport-Guided Source-Free Adaptation for Face Anti-Spoofing
Zhuowei Li, Tianchen Zhao, Xiang Xu, Zheng Zhang, Zhihua Li, Xuanbai Chen, Qin Zhang, Alessandro Bergamo, Anil K. Jain, Yifan Xing; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24351-24363
Developing a face anti-spoofing model that meets the security requirements of clients worldwide is challenging due to the domain gap between training datasets and the diverse end-user test data. Moreover, for security and privacy reasons, it is undesirable for clients to share a large amount of their face data with service providers. In this work, we introduce a novel method in which the face anti-spoofing model can be adapted by the client itself to a target domain at test time using only a small sample of data while keeping model parameters and training data inaccessible to the client. Specifically, we develop a prototype-based base model and an optimal transport-guided adaptor that enables adaptation in either a lightweight training or training-free fashion, without updating the base model's parameters. Furthermore, we propose geodesic mixup, an optimal transport-based synthesis method that generates augmented training data along the geodesic path between source prototypes and the target data distribution. This allows training a lightweight classifier to effectively adapt to target-specific characteristics while retaining essential knowledge learned from the source domain. In cross-domain and cross-attack settings, compared with recent methods, our method achieves average relative improvements of 19.17% in HTER and 8.58% in AUC, respectively.

****

USP-Gaussian: Unifying Spike-based Image Reconstruction, Pose Correction and Gaussian Splatting
Kang Chen, Jiyuan Zhang, Zecheng Hao, Yajing Zheng, Tiejun Huang, Zhaofei Yu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16609-16618
Spike camera, as an innovative type of neuromorphic camera that captures scenes with 0-1 bit stream at 40 kHz, is increasingly being employed for the novel view synthesis task building on the techniques such as Neural Radiance Fields (NeRF) and 3D Gaussian Splatting (3DGS). Previous spike-based approaches typically follow a three-stage pipeline: I. Spike-to-image reconstruction based on established algorithms. II. Camera poses estimation. III. Novel view synthesis. However, the cascading framework suffers from substantial cumulative errors, i.e., the quality of the initially reconstructed images will impact pose estimation, ultimately limiting the fidelity of the 3D reconstruction. To address this limitation, we propose a synergistic optimization framework USP-Gaussian, which unifies spike-to-image reconstruction, pose correction, and gaussian splatting into an end-to

-end pipeline. Leveraging the multi-view consistency afforded by 3DGS and the mo
tion capture capability of the spike camera, our framework enables iterative opt
imization between the spike-to-image reconstruction network and 3DGS. Experiment
s on synthetic datasets demonstrate that our method surpasses previous approache
s by effectively eliminating cascading errors. Moreover, in real-world scenarios
, our method achieves robust 3D reconstruction benefiting from the integration o
f pose optimization. Our code, data, and trained models are available at https:/
/github.com/chenkang455/USP-Gaussian.
****

Robust 3D Shape Reconstruction in Zero-Shot from a Single Image in the Wild
Junhyeong Cho, Kim Youwang, Hunmin Yang, Tae-Hyun Oh; Proceedings of the Compute
r Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22786-22798
Recent monocular 3D shape reconstruction methods have shown promising zero-shot
results on object-segmented images without any occlusions. However, their effect
iveness is significantly compromised in real-world conditions, due to imperfect
object segmentation by off-the-shelf models and the prevalence of occlusions. To
 effectively address these issues, we propose a unified regression model that in
tegrates segmentation and reconstruction, specifically designed for occlusion-aw
are 3D shape reconstruction. To facilitate its reconstruction in the wild, we al
so introduce a scalable data synthesis pipeline that simulates a wide range of v
ariations in objects, occluders, and backgrounds. Training on our synthetic data
 enables the proposed model to achieve state-of-the-art zero-shot results on rea
l-world images, using significantly fewer parameters than competing approaches.
****

BOE-ViT: Boosting Orientation Estimation with Equivariance in Self-Supervised 3D
 Subtomogram Alignment
Runmin Jiang, Jackson Daggett, Shriya Pingulkar, Yizhou Zhao, Priyanshu Dhingra,
 Daniel Brown, Qifeng Wu, Xiangrui Zeng, Xingjian Li, Min Xu; Proceedings of the
 Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29352-2936
2
Subtomogram alignment is a critical task in cryo-electron tomography (cryo-ET) a
nalysis, essential for achieving high-resolution reconstructions of macromolecul
ar complexes. However, learning effective positional representations remains cha
llenging due to limited labels and high noise levels inherent in cryo-ET data. I
n this work, we address this challenge by proposing a self-supervised learning a
pproach that leverages intrinsic geometric transformations as implicit superviso
ry signals, enabling robust representation learning despite data scarcity. We in
troduce BOE-ViT, the first Vision Transformer (ViT) framework for 3D subtomogram
 alignment. Recognizing that traditional ViTs lack equivariance and are therefor
e suboptimal for orientation estimation, we enhance the model with two innovativ
e modules that introduce equivariance include 1) the Polyshift module for improv
ed shift estimation and 2) Multi-Axis Rotation Encoding (MARE) for enhanced rota
tion estimation. Experimental results demonstrate that BOE-ViT significantly out
performs state-of-the-art methods. Notably, at SNR 0.01 dataset, our approach ac
hieves a 77.3% reduction in rotation estimation error and a 62.5% reduction in t
ranslation estimation error, effectively overcoming the challenges in cryo-ET su
btomogram alignment.
****

Holmes-VAU: Towards Long-term Video Anomaly Understanding at Any Granularity
Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao,
Shanjun Zhang, Li Yu, Nong Sang; Proceedings of the Computer Vision and Pattern
Recognition Conference (CVPR), 2025, pp. 13843-13853
How can we enable models to comprehend video anomalies occurring over varying te
mporal scales and contexts?Traditional Video Anomaly Understanding (VAU) methods
 focus on frame-level anomaly prediction, often missing the interpretability of
complex and diverse real-world anomalies. Recent multimodal approaches leverage
visual and textual data but lack hierarchical annotations that capture both shor
t-term and long-term anomalies.To address this challenge, we introduce HIVAU-70k
, a large-scale benchmark for hierarchical video anomaly understanding across an
y granularity. We develop a semi-automated annotation engine that efficiently sc

ales high-quality annotations by combining manual video segmentation with recurs ive free-text annotation using large language models (LLMs). This results in ove r 70,000 multi-granular annotations organized at clip-level, event-level, and vi deo-level segments.For efficient anomaly detection in long videos, we propose th e Anomaly-focused Temporal Sampler (ATS). ATS integrates an anomaly scorer with a density-aware sampler to adaptively select frames based on anomaly scores, ens uring that the multimodal LLM concentrates on anomaly-rich regions, which signif icantly enhances both efficiency and accuracy.Extensive experiments demonstrate that our hierarchical instruction data markedly improves anomaly comprehension. The integrated ATS and visual-language model outperform traditional methods in p rocessing long videos.Our benchmark and model will be publicly available.
****

Adventurer: Optimizing Vision Mamba Architecture Designs for Efficiency
Feng Wang, Timing Yang, Yaodong Yu, Sucheng Ren, Guoyizhe Wei, Angtian Wang, Wei Shao, Yuyin Zhou, Alan Yuille, Cihang Xie; Proceedings of the Computer Vision a nd Pattern Recognition Conference (CVPR), 2025, pp. 30157-30166
In this work, we introduce the Adventurer series models where we treat images as sequences of patch tokens and employ uni-directional language models to learn v isual representations. This modeling paradigm allows us to process images in a r ecurrent formulation with linear complexity relative to the sequence length, whi ch can effectively address the memory and computation explosion issues posed by high-resolution and fine-grained images. In detail, we introduce two simple desi gns that seamlessly integrate image inputs into the causal inference framework: a global pooling token placed at the beginning of the sequence and a flipping op eration between every two layers. Extensive empirical studies highlight that com pared with the existing plain architectures such as DeiT and Vim, Adventurer off ers an optimal efficiency-accuracy trade-off. For example, our Adventurer-Base a ttains a competitive test accuracy of 84.3% on the standard ImageNet-1k benchmar k with 216 images/s training throughput, which is 3.8x and 6.2x faster than Vim and DeiT to achieve the same result. As Adventurer offers great computation and memory efficiency and allows scaling with linear complexity, we hope this archit ecture can benefit future explorations in modeling long sequences for high-resol ution or fine-grained images.
****

Beyond Local Sharpness: Communication-Efficient Global Sharpness-aware Minimizat ion for Federated Learning
Debora Caldarola, Pietro Cagnasso, Barbara Caputo, Marco Ciccone; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25187-25197
Federated learning (FL) enables collaborative model training with privacy preser vation. Data heterogeneity across edge devices (clients) can cause models to con verge to sharp minima, negatively impacting generalization and robustness. Recen t approaches use client-side sharpness-aware minimization (SAM) to encourage fla tter minima, but the discrepancy between local and global loss landscapes often undermines their effectiveness, as optimizing for local sharpness does not ensur e global flatness. This work introduces FedGloSS (Federated Global Server-side S harpness), a novel FL approach that prioritizes the optimization of global sharp ness on the server, using SAM. To reduce communication overhead, FedGloSS clever ly approximates sharpness using the previous global gradient, eliminating the ne ed for additional client communication. Our extensive evaluations demonstrate th at FedGloSS consistently reaches flatter minima and better performance compared to state-of-the-art FL methods across various federated vision benchmarks. Code available at https://github.com/pietrocagnasso/fedgloss.
****

ALIEN: Implicit Neural Representations for Human Motion Prediction under Arbitra ry Latency
Dong Wei, Xiaoning Sun, Xizhan Gao, Shengxiang Hu, Huaijiang Sun; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1861-1 870
We investigate a new task in human motion prediction, which aims to forecast fut

ure body poses from historically observed sequences while accounting for arbitrary latency. This differs from existing works that assume an ideal scenario where future motions can be "instantaneously" predicted, thereby neglecting time delays caused by network transmission and algorithm execution. Addressing this task requires tackling two key challenges: The length of latency period can vary significantly across samples; the prediction model must be efficient. In this paper, we propose ALIEN, which treats the motion as a continuous function parameterized by a neural network, enabling predictions under any latency condition. By incorporating Mamba-like linear attention as a hyper-network and designing subsequent low-rank modulation, ALIEN efficiently learns a set of implicit neural representation weights from the observed motion to encode instance-specific information. Additionally, our model integrates the primary motion prediction task with an extra-designed variable-delay pose reconstruction task in a unified multi-task learning framework, enhancing its ability to capture richer motion patterns. Extensive experiments demonstrate that our approach outperforms state-of-the-art baselines adapted for our new task, while maintaining competitive performance in traditional prediction setting.

****

Parameterized Blur Kernel Prior Learning for Local Motion Deblurring

Zhenxuan Fang, Fangfang Wu, Tao Huang, Le Dong, Weisheng Dong, Xin Li, Guangming Shi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23006-23015

Unlike global motion blur, Local Motion Deblurring (LMD) presents a more complex challenge, as it requires precise restoration of blurry regions while preserving the sharpness of the background. Existing LMD methods rely on manually annotated blur masks and often overlook the blur kernel's characteristics, which are crucial for accurate restoration. To address these limitations, we propose a novel parameterized motion kernel modeling approach that defines the motion blur kernel with three key parameters: length, angle, and curvature. We then use networks to estimate these kernel parameters, significantly improving the accuracy of blur kernel estimation. To effectively learn the motion blur representation, we incorporate a shared memory bank that stores blur prior information. Additionally, we introduce a dual-branch deblurring network: one branch leverages Mamba to capture long-range dependencies, while the other uses a mask-guided CNN focused on refining the local blurry regions. By fully utilizing the estimated bur prior information, our approach greatly enhances deblurring outcomes. Experimental results show that our method achieves state-of-the-art performance both quantitatively and visually, with a substantial reduction in computational complexity.

****

QuartDepth: Post-Training Quantization for Real-Time Depth Estimation on the Edge

Xuan Shen, Weize Ma, Jing Liu, Changdi Yang, Rui Ding, Quanyi Wang, Henghui Ding, Wei Niu, Yanzhi Wang, Pu Zhao, Jun Lin, Jiuxiang Gu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11448-11460

Monocular Depth Estimation (MDE) has emerged as a pivotal task in computer vision, supporting numerous real-world applications. However, deploying accurate depth estimation models on resource-limited edge devices, especially Application-Specific Integrated Circuits (ASICs), is challenging due to the high computational and memory demands. Recent advancements in foundational depth estimation deliver impressive results but further amplify the difficulty of deployment on ASICs. To address this, we propose QuartDepth which adopts post-training quantization to quantize MDE models with hardware accelerations for ASICs. Our approach involves quantizing both weights and activations to 4-bit precision, reducing the model size and computation cost. To mitigate the performance degradation, we introduce activation polishing and compensation algorithm applied before and after activation quantization, as well as a weight reconstruction method for minimizing errors in weight quantization. Furthermore, we design a flexible and programmable hardware accelerator by supporting kernel fusion and customized instruction programmability, enhancing throughput and efficiency. Experimental results demonstrate that our framework achieves competitive accuracy while enabling fast inference

and higher energy efficiency on ASICs, bridging the gap between high-performance depth estimation and practical edge-device applicability. Code: https://github.com/shawnricecake/quart-depth

****

ReWind: Understanding Long Videos with Instructed Learnable Memory

Anxhelo Diko, Tinghuai Wang, Wassim Swaileh, Shiyan Sun, Ioannis Patras; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13734-13743

Vision-Language Models (VLMs) are crucial for real-world applications that require understanding textual and visual information. However, existing VLMs face multiple challenges in processing long videos, including computational inefficiency, memory limitations, and difficulties maintaining coherent understanding across extended sequences. These issues stem partly from the quadratic scaling of self-attention w.r.t. number of tokens but also encompass broader challenges in temporal reasoning and information integration over long sequences. To address these challenges, we introduce ReWind, a novel two-stage framework for long video understanding. In the first stage, ReWind maintains a dynamic memory that stores and updates instruction-relevant visual information as the video unfolds.Memory updates leverage novel read and write mechanisms utilizing learnable queries and cross-attentions between memory contents and the input stream. This approach maintains low memory requirements as the cross-attention layers scale linearly w.r.t. number of tokens. In the second stage, the memory content guides the selection of a few relevant frames, represented at high spatial resolution, which are combined with the memory contents and fed into an LLM to generate the final answer. We empirically demonstrate ReWind's superiority in visual question answering (VQA) and temporal grounding tasks, surpassing previous methods on long video benchmarks. Notably, ReWind achieves a +13% score gain and a +12% accuracy improvement on the MovieChat-1K VQA dataset and an +8% mIoU increase on Charades-STA for temporal grounding.

****

Sufficient Invariant Learning for Distribution Shift

Taero Kim, Subeen Park, Sungjun Lim, Yonghan Jung, Krikamol Muandet, Kyungwoo Song; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4958-4967

Learning robust models under distribution shifts between training and test datasets is a fundamental challenge in machine learning. While learning invariant features across environments is a popular approach, it often assumes that these features are fully observed in both training and test sets--a condition frequently violated in practice. When models rely on invariant features absent in the test set, their robustness in new environments can deteriorate. To tackle this problem, we introduce a novel learning principle called the Sufficient Invariant Learning (SIL) framework, which focuses on learning a sufficient subset of invariant features rather than relying on a single feature. After demonstrating the limitation of existing invariant learning methods, we propose a new algorithm, Adaptive Sharpness-aware Group Distributionally Robust Optimization (ASGDRO), to learn diverse invariant features by seeking common flat minima across the environments. We theoretically demonstrate that finding a common flat minima enables robust predictions based on diverse invariant features. Empirical evaluations on multiple datasets, including our new benchmark, confirm ASGDRO's robustness against distribution shifts, highlighting the limitations of existing methods.

****

DirectTriGS: Triplane-based Gaussian Splatting Field Representation for 3D Generation

Xiaoliang Ju, Hongsheng Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16229-16239

We present DirectTriGS, a novel framework designed for 3D object generation with Gaussian Splatting (GS). GS-based rendering for 3D content has gained considerable attention recently. However, there has been limited exploration in directly generating 3D Gaussians compared to traditional generative modeling approaches. The main challenge lies in the complex data structure of GS represented by discr

ete point clouds with multiple channels.To overcome this challenge, we propose e mploying the triplane representation, which allows us to represent Gaussian Spla tting as an image-like continuous field. This representation effectively encodes both the geometry and texture information, enabling smooth transformation back to Gaussian point clouds and rendering into images by a TriRenderer, with only 2 D supervisions. The proposed TriRenderer is fully differentiable, so that the re ndering loss can supervise both texture and geometry encoding. Furthermore, the triplane representation can be compressed using a Variational Autoencoder (VAE), which can subsequently be utilized in latent diffusion to generate 3D objects.T he experiments demonstrate that the proposed generation framework can produce hi gh-quality 3D object geometry and rendering results.

****

Domain Generalization in CLIP via Learning with Diverse Text Prompts

Changsong Wen, Zelin Peng, Yu Huang, Xiaokang Yang, Wei Shen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9559-9569

Domain generalization (DG) aims to train a model on source domains that can gene ralize well to unseen domains. Recent advances in Vision-Language Models (VLMs), such as CLIP, exhibit remarkable generalization capabilities across a wide rang e of data distributions, benefiting tasks like DG. However, CLIP is pre-trained by aligning images with their descriptions, which inevitably captures domain-spe cific details. Moreover, adapting CLIP to source domains with limited feature di versity introduces bias. These limitations hinder the model's ability to general ize across domains. In this paper, we propose a new DG approach by learning with diverse text prompts. These text prompts incorporate varied contexts to imitate different domains, enabling DG model to learn domain-invariant features. The te xt prompts guide DG model learning in three aspects: feature suppression, which uses these prompts to identify domain-sensitive features and suppress them; feat ure consistency, which ensures the model's features are robust to domain variati ons imitated by the diverse prompts; and feature diversification, which diversif ies features based on the prompts to mitigate bias. Experimental results show th at our approach improves domain generalization performance on five datasets on t he DomainBed benchmark, achieving state-of-the-art results.

****

Scene4U: Hierarchical Layered 3D Scene Reconstruction from Single Panoramic Imag e for Your Immerse Exploration

Zilong Huang, Jun He, Junyan Ye, Lihan Jiang, Weijia Li, Yiping Chen, Ting Han; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 20 25, pp. 26723-26733

The reconstruction of immersive and realistic 3D scenes holds significant practi cal importance in various fields of computer vision and computer graphics. Typic ally, immersive and realistic scenes should be free from obstructions by dynamic objects, maintain global texture consistency, and allow for unrestricted explor ation. The current mainstream methods for image-driven scene construction involv es iteratively refining the initial image using a moving virtual camera to gener ate the scene. However, previous methods struggle with visual discontinuities du e to global texture inconsistencies under varying camera poses, and they frequen tly exhibit scene voids caused by foreground-background occlusions. To this end, we propose a novel layered 3D scene reconstruction framework from panoramic ima ge, named Scene4U. Specifically, Scene4U integrates an open-vocabulary segmentat ion model with a large language model to decompose a real panorama into multiple layers. Then, we employs a layered repair module based on diffusion model to re store occluded regions using visual cues and depth information, generating a hie rarchical representation of the scene. The multi-layer panorama is then initiali zed as a 3D Gaussian Splatting representation, followed by layered optimization, which ultimately produces an immersive 3D scene with semantic and structural co nsistency that supports free exploration. Our Scene4U outperforms state-of-the-a rt method, improving by 24.24% in LPIPS and 24.40% in BRISQUE, while also achiev ing the fastest training speed. Additionally, to demonstrate the robustness of S cene4U and allow users to experience immersive scenes from various landmarks, we build WorldVista3D dataset for 3D scene reconstruction, which contains panorami

c images of globally renowned sites. The implementation code and dataset will be made publicly available.
****

Make-It-Animatable: An Efficient Framework for Authoring Animation-Ready 3D Characters
Zhiyang Guo, Jinxu Xiang, Kai Ma, Wengang Zhou, Houqiang Li, Ran Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10783-10792
3D characters are essential to modern creative industries, but making them animatable often demands extensive manual work in tasks like rigging and skinning. Existing automatic rigging tools face several limitations, including the necessity for manual annotations, rigid skeleton topologies, and limited generalization across diverse shapes and poses. An alternative approach generates animatable avatars pre-bound to a rigged template mesh. However, this method often lacks flexibility and is typically limited to realistic human shapes. To address these issues, we present Make-It-Animatable, a novel data-driven method to make any 3D humanoid model ready for character animation in less than one second, regardless of its shapes and poses. Our unified framework generates high-quality blend weights, bones, and pose transformations. By incorporating a particle-based shape autoencoder, our approach supports various 3D representations, including meshes and 3D Gaussian splats. Additionally, we employ a coarse-to-fine representation and a structure-aware modeling strategy to ensure both accuracy and robustness, even for characters with non-standard skeleton structures. We conducted extensive experiments to validate our framework's effectiveness. Compared to existing methods, our approach demonstrates significant improvements in both quality and speed. More demos and code are available at https://jasongzy.github.io/Make-It-Animatable/.
****

IterIS: Iterative Inference-Solving Alignment for LoRA Merging
Hongxu Chen, Zhen Wang, Runshi Li, Bowei Zhu, Long Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4829-4838
Low-rank adaptations (LoRA) are widely used to fine-tune large models across various domains for specific downstream tasks. While task-specific LoRAs are often available, concerns about data privacy and intellectual property can restrict access to training data, limiting the acquisition of a multi-task model through gradient-based training. In response, LoRA merging presents an effective solution by combining multiple LoRAs into a unified adapter while maintaining data privacy. Prior works on LoRA merging primarily frame it as an optimization problem, yet these approaches face several limitations, including the rough assumption about input features utilized in optimization, massive sample requirements, and the unbalanced optimization objective. These limitations can significantly degrade performance. To address these, we propose a novel optimization-based method, named IterIS: 1) We formulate LoRA merging as an advanced optimization problem to mitigate the rough assumption. Additionally, we employ an iterative inference-solving framework in our algorithm. It can progressively refine the optimization objective for improved performance. 2) We introduce an efficient regularization term to reduce the need for massive sample requirements (requiring only 1-5% of the unlabeled samples compared to prior methods). 3) We utilize adaptive weights in the optimization objective to mitigate potential unbalances in LoRA merging process. Our method demonstrates significant improvements over multiple baselines and state-of-the-art methods in composing tasks for text-to-image diffusion, vision-language models, and large language models. Furthermore, our layer-wise algorithm can achieve convergence with minimal steps, ensuring efficiency in both memory and computation.
****

ACAttack: Adaptive Cross Attacking RGB-T Tracker via Multi-Modal Response Decoupling
Xinyu Xiang, Qinglong Yan, Hao Zhang, Jiayi Ma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22099-22108
The research on adversarial attacks against trackers primarily concentrates on t

he RGB modality, whereas the methodology for attacking RGB-T multi-modal trackers has seldom been explored so far. This work represents an innovative attempt to develop an adaptive cross attack framework via multi-modal response decoupling, generating multi-modal adversarial patches to evade RGB-T trackers. Specifically, a modal-aware adaptive attack strategy is introduced to weaken the modality with high common information contribution alternately and iteratively, achieving the modal decoupling attack. In order to perturb the judgment of the modal balance mechanism in the tracker, we design a modal disturbance loss to increase the distance of the response map of the single-modal adversarial samples in the tracker. Besides, we also propose a novel spatio-temporal joint attack loss to progressively deteriorate the tracker's perception of the target. Moreover, the design of the shared adversarial shape enables the generated multi-modal adversarial patches to be readily deployed in real-world scenarios, effectively reducing the interference of the patch posting process on the shape attack of the infrared adversarial layer. Extensive digital and physical domain experiments demonstrate the effectiveness of our multi-modal adversarial patch attack. Our code is available at https://github.com/Xinyu-Xiang/ACAttack.

****

DeCafNet: Delegate and Conquer for Efficient Temporal Grounding in Long Videos

Zijia Lu, A S M Iftekhar, Gaurav Mittal, Tianjian Meng, Xiawei Wang, Cheng Zhao, Rohith Kukkala, Ehsan Elhamifar, Mei Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24066-24076

Long Video Temporal Grounding (LVTG) aims at identifying specific moments within lengthy videos based on user-provided text queries for effective content retrieval. The approach taken by existing methods of dividing video into clips and processing each clip via a full-scale expert encoder is challenging to scale due to prohibitive computational costs of processing a large number of clips in long videos. To address this issue, we introduce DeCafNet, an approach employing "delegate-and-conquer" strategy to achieve computation efficiency without sacrificing grounding performance. DeCafNet introduces a sidekick encoder that performs dense feature extraction over all video clips in a resource-efficient manner, while generating a saliency map to identify the most relevant clips for full processing by the expert encoder. To effectively leverage features from sidekick and expert encoders that exist at different temporal resolutions, we introduce DeCaf-Grounder, which unifies and refines them via query-aware temporal aggregation and multi-scale temporal refinement for accurate grounding. Experiments on two LTVG benchmark datasets demonstrate that DeCafNet reduces computation by up to 47% while still outperforming existing methods, establishing a new state-of-the-art for LTVG in terms of both efficiency and performance.

****

Efficient ANN-Guided Distillation: Aligning Rate-based Features of Spiking Neural Networks through Hybrid Block-wise Replacement

Shu Yang, Chengting Yu, Lei Liu, Hanzhi Ma, Aili Wang, Erping Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10025-10035

Spiking Neural Networks (SNNs) have garnered considerable attention as a potential alternative to Artificial Neural Networks (ANNs). Recent studies have highlighted SNNs' potential on large-scale datasets. For SNN training, two main approaches exist: direct training and ANN-to-SNN (ANN2SNN) conversion. To fully leverage existing ANN models in guiding SNN learning, either direct ANN-to-SNN conversion or ANN-SNN distillation training can be employed. In this paper, we propose an ANN-SNN distillation framework from the ANN-to-SNN perspective, designed with a block-wise replacement strategy for ANN-guided learning. By generating intermediate hybrid models that progressively align SNN feature spaces to those of ANN through rate-based features, our framework naturally incorporates rate-based backpropagation as a training method. Our approach achieves results comparable to or better than state-of-the-art SNN distillation methods, showing both training and learning efficiency.

****

PrEditor3D: Fast and Precise 3D Shape Editing

Ziya Erkoç, Can Gümeli, Chaoyang Wang, Matthias Nießner, Angela Dai, Peter Wonka , Hsin-Ying Lee, Peiye Zhuang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 640-649

We propose a training-free approach to 3D editing that enables the editing of a single shape and the reconstruction of a mesh within a few minutes. Leveraging 4-view images, user-guided text prompts, and rough 2D masks, our method produces an edited 3D mesh that aligns with the prompt. For this, our approach performs synchronized multi-view image editing in 2D. However, targeted regions to be edited are ambiguous due to projection from 3D to 2D. To ensure precise editing only in intended regions, we develop a 3D segmentation pipeline that detects edited areas in 3D space. Additionally, we introduce a merging algorithm to seamlessly integrate edited 3D regions with original input. Extensive experiments demonstrate the superiority of our method over previous approaches, enabling fast, high-quality editing while preserving unintended regions.

****

Subspace Constraint and Contribution Estimation for Heterogeneous Federated Learning

Xiangtao Zhang, Sheng Li, Ao Li, Yipeng Liu, Fan Zhang, Ce Zhu, Le Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20632-20642

Heterogeneous Federated Learning (HFL) has received widespread attention due to its adaptability to different models and data. The HFL approach utilizing auxiliary models for knowledge transfer enhances flexibility. However, existing frameworks face the challenges of aggregation bias and local overfitting. To address these issues, we propose FedSCE. It reduces the degree of freedom of update and improves generalization performance by limiting the specific layer of local model update to the local subspace. The subspace is dynamically updated to ensure coverage of the latest model update trajectory. Additionally, FedSCE evaluates client contributions based on the update distance of the auxiliary model in feature space and parameter space, achieving adaptive weighted aggregation. We validate our approach in both feature-skewed and label-skewed scenarios, demonstrating that on Office10, our method exceeds the best baseline by 3.87%. The code will be available at https://github.com/AVC2-UESTC/FedSCE.git.

****

HoGS: Unified Near and Far Object Reconstruction via Homogeneous Gaussian Splatting

Xinpeng Liu, Zeyi Huang, Fumio Okura, Yasuyuki Matsushita; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26714-26722

Novel view synthesis has demonstrated impressive progress recently, with 3D Gaussian splatting (3DGS) offering efficient training time and photorealistic real-time rendering. However, reliance on Cartesian coordinates limits 3DGS's performance on distant objects, which is important for reconstructing unbounded outdoor environments. We found that, despite its ultimate simplicity, using homogeneous coordinates, a concept on the projective geometry, for the 3DGS pipeline remarkably improves the rendering accuracies of distant objects. We therefore propose Homogeneous Gaussian Splatting (HoGS) incorporating homogeneous coordinates into the 3DGS framework, providing a unified representation for enhancing near and distant objects. HoGS effectively manages both expansive spatial positions and scales particularly in outdoor unbounded environments by adopting projective geometry principles. Experiments show that HoGS significantly enhances accuracy in reconstructing distant objects while maintaining high-quality rendering of nearby objects, along with fast training speed and real-time rendering capability. Our implementation will be released upon acceptance. Our implementations are available on our project page https://kh129.github.io/hogs/.

****

SmartEraser: Remove Anything from Images using Masked-Region Guidance

Longtao Jiang, Zhendong Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Lei Shi, Dong Chen, Houqiang Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24452-24462

Object removal has so far been dominated by the mask-and-inpaint paradigm, where

the masked region is excluded from the input, leaving models relying on unmaske
d areas to inpaint the missing region. However, this approach lacks contextual i
nformation for the masked area, often resulting in unstable performance. In this
 work, we introduce SmartEraser, built with a new removing paradigm called Maske
d-Region Guidance. This paradigm retains the masked region in the input, using i
t as guidance for the removal process. It offers several distinct advantages: (a
) it guides the model to accurately identify the object to be removed, preventin
g its regeneration in the output; (b) since the user mask often extends beyond t
he object itself, it aids in preserving the surrounding context in the final res
ult. Leveraging this new paradigm, we present Syn4Removal, a large-scale object
 removal dataset, where instance segmentation data is used to copy and paste obje
cts onto images as removal targets, with the original images serving as ground t
ruths. Experimental results demonstrate that SmartEraser significantly outperfor
ms existing methods, achieving superior performance in object removal, especiall
y in complex scenes with intricate compositions.
****

ComRoPE: Scalable and Robust Rotary Position Embedding Parameterized by Trainabl
e Commuting Angle Matrices

Hao Yu, Tangyu Jiang, Shuning Jia, Shannan Yan, Shunning Liu, Haolong Qian, Guan
ghao Li, Shuting Dong, Chun Yuan; Proceedings of the Computer Vision and Pattern
 Recognition Conference (CVPR), 2025, pp. 4508-4517

The Transformer architecture has revolutionized various fields since it was prop
osed, where positional encoding plays an essential role in effectively capturing
 sequential order and context. Therefore, Rotary Positional Encoding (RoPE) was
proposed to alleviate these issues, which integrates positional information by r
otating the embeddings in the attention mechanism. However, RoPE utilizes manual
ly defined rotation matrices, a design choice that favors computational efficien
cy but limits the model's flexibility and adaptability. In this work, we propose
 ComRoPE, which generalizes RoPE by defining it in terms of trainable commuting
angle matrices. Specifically, we demonstrate that pairwise commutativity of thes
e matrices is essential for RoPE to achieve scalability and positional robustnes
s. We formally define the RoPE Equation, which is an essential condition that en
sures consistent performance with position offsets. Based on the theoretical ana
lysis, we present two types of trainable commuting angle matrices as sufficient
solutions to the RoPE equation, which significantly improve performance, surpass
ing the current state-of-the-art method by 1.6% at training resolution and 2.9%
at higher resolution on the ImageNet-1K dataset. Furthermore, our framework show
s versatility in generalizing to existing RoPE formulations and offering new ins
ights for future positional encoding research. To ensure reproducibility, the so
urce code and instructions are available at https://github.com/Longin-Yu/ComRoPE
****

Sample- and Parameter-Efficient Auto-Regressive Image Models

Elad Amrani, Leonid Karlinsky, Alex Bronstein; Proceedings of the Computer Visio
n and Pattern Recognition Conference (CVPR), 2025, pp. 30127-30136

We introduce XTRA, a vision model pre-trained with a novel auto-regressive objec
tive that significantly enhances both sample and parameter efficiency compared t
o previous auto-regressive image models. Unlike contrastive or masked image mode
ling methods, which have not been demonstrated as having consistent scaling beha
vior on unbalanced internet data,auto-regressive vision models exhibit scalable
and promising performance as model and dataset size increase. In contrast to sta
ndard auto-regressive models, XTRA employs a Block Causal Mask, where each Block
 represents k x k tokens rather than relying on a standard causal mask. By recon
structing pixel values block by block, XTRA captures higher-level structural pat
terns over larger image regions. Predicting on blocks allows the model to learn
relationships across broader areas of pixels, enabling more abstract and semanti
cally meaningful representations than traditional next-token prediction.This sim
ple modification yields two key results. First, XTRA is sample-efficient. Despit
e being trained on 152x fewer samples (13.1M vs. 2B), XTRA ViT-H/14 surpasses th
e top-1 average accuracy of the previous state-of-the-art auto-regressive model
across 15 diverse image recognition benchmarks. Second, XTRA is parameter-effici

ent. Compared to auto-regressive models trained on ImageNet-1k, XTRA ViT-B/16 ou tperforms in linear and attentive probing tasks, using 7-16x fewer parameters (8 5M vs. 1.36B/0.63B).
****

Robust Audio-Visual Segmentation via Audio-Guided Visual Convergent Alignment
Chen Liu, Peike Li, Liying Yang, Dadong Wang, Lincheng Li, Xin Yu; Proceedings o f the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28922 -28931
Accurately localizing audible objects based on audio-visual cues is the core obj ective of audio-visual segmentation. Most previous methods emphasize spatial or temporal multi-modal modeling, yet overlook challenges from ambiguous audio-visu al correspondences--such as nearby visually similar but acoustically different o bjects and frequent shifts in objects' sounding status. Consequently, they may s truggle to reliably correlate audio and visual cues, leading to over- or under-s egmentation. To address these limitations, we propose a novel framework with two primary components: an audio-guided modality alignment (AMA) module and an unce rtainty estimation (UE) module. Instead of indiscriminately correlating audio-vi sual cues through a global attention mechanism, AMA performs audio-visual intera ctions within multiple groups and consolidates group features into compact repre sentations based on their responsiveness to audio cues, effectively directing th e model's attention to audio-relevant areas. Leveraging contrastive learning, A MA further distinguishes sounding regions from silent areas by treating features with strong audio responses as positive samples and weaker responses as negativ es. Additionally, UE integrates spatial and temporal information to identify hig h-uncertainty regions caused by frequent changes in sound state, reducing predic tion errors by lowering confidence in these areas. Experimental results demonstr ate that our approach achieves superior accuracy compared to existing state-of-t he-art methods, particularly in challenging scenarios where traditional approach es struggle to maintain reliable segmentation.
****

LOCORE: Image Re-ranking with Long-Context Sequence Modeling
Zilin Xiao, Pavel Suma, Ayush Sachdeva, Hao-Jen Wang, Giorgos Kordopatis-Zilos, Giorgos Tolias, Vicente Ordonez; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9580-9590
We introduce LOCORE, Long-Context Re-ranker, a model that takes as input local d escriptors corresponding to an image query and a list of gallery images and outp uts similarity scores between the query and each gallery image. This model is us ed for image retrieval, where typically a first ranking is performed with an eff icient similarity measure, and then a shortlist of top-ranked images is re-ranke d based on a more fine-grained similarity measure. Compared to existing methods that perform pair-wise similarity estimation with local descriptors or list-wise re-ranking with global descriptors, LOCORE is the first method to perform list-wise re-ranking with local descriptors. To achieve this, we leverage efficient l ong-context sequence models to effectively capture the dependencies between quer y and gallery images at the local-descriptor level. During testing, we process l ong shortlists with a sliding window strategy that is tailored to overcome the c ontext size limitations of sequence models. Our approach achieves superior perfo rmance compared with other re-rankers on established image retrieval benchmarks of landmarks (\mathcal R Oxf and \mathcal R Par), products (SOP), fashion items (In-Shop), and bird species (CUB-200) while having comparable latency to the pai r-wise local descriptor re-rankers.
****

NeRFPrior: Learning Neural Radiance Field as a Prior for Indoor Scene Reconstruc tion
Wenyuan Zhang, Emily Yue-ting Jia, Junsheng Zhou, Baorui Ma, Kanle Shi, Yu-Shen Liu, Zhizhong Han; Proceedings of the Computer Vision and Pattern Recognition Co nference (CVPR), 2025, pp. 11317-11327
Recently, it has shown that priors are vital for neural implicit functions to re construct high-quality surfaces from multi-view RGB images. However, current pri ors require large-scale pre-training, and merely provide geometric clues without

considering the importance of color. In this paper, we present NeRFPrior, which adopts a neural radiance field as a prior to learn signed distance fields using volume rendering for surface reconstruction. Our NeRF prior can provide both geometric and color clues, and also get trained fast under the same scene without additional data. Based on the NeRF prior, we are enabled to learn a signed distance function (SDF) by explicitly imposing a multi-view consistency constraint on each ray intersection for surface inference. Specifically, at each ray intersection, we use the density in the prior as a coarse geometry estimation, while using the color near the surface as a clue to check its visibility from another view angle. For the textureless areas where the multi-view consistency constraint does not work well, we further introduce a depth consistency loss with confidence weights to infer the SDF. Our experimental results outperform the state-of-the-art methods under the widely used benchmarks. The source code will be publicly available.

****

BiLoRA: Almost-Orthogonal Parameter Spaces for Continual Learning

Hao Zhu, Yifei Zhang, Junhao Dong, Piotr Koniusz; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25613-25622

Continual learning requires models to learn tasks sequentially while maintaining a delicate balance between stability (retaining knowledge of previous tasks) and plasticity (adapting to new tasks). A key challenge is preventing interference between tasks - where learning new tasks degrades performance on previously learned ones. Recent approaches have leveraged parameter-efficient fine-tuning (PEFT) methods, which adapt pre-trained models by injecting a small number of learnable parameters. However, existing PEFT-based continual learning methods like InfLoRA face fundamental limitations: they rely on complex optimization procedures to learn orthogonal task-specific spaces, and finding such spaces becomes increasingly difficult as tasks accumulate. We propose a novel bilinear reformulation that fundamentally reimagines task separation through fixed orthogonal bases. Our key insight is that by expanding the parameter space quadratically through two fixed bases, we can achieve "almost orthogonal" task subspaces probabilistically, eliminating the need for explicit interference elimination procedures. We provide theoretical guarantees that this approach reduces the probability of task interference from \bigO((k/d)^2) to \bigO((k/d^2)^2), ensuring reliable task separation without complex optimization. Through extensive experiments on ImageNet-R, CIFAR100, and DomainNet, we validate our theoretical bounds and demonstrate state-of-the-art performance with reduced parameter count.

****

Vid2Sim: Generalizable, Video-based Reconstruction of Appearance, Geometry and Physics for Mesh-free Simulation

Chuhao Chen, Zhiyang Dou, Chen Wang, Yiming Huang, Anjun Chen, Qiao Feng, Jiatao Gu, Lingjie Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26545-26555

Faithfully reconstructing textured shapes and physical properties from videos presents an intriguing yet challenging problem. Significant efforts have been dedicated to advancing such a system identification problem in this area. Previous methods often rely on heavy optimization pipelines with a differentiable simulator and renderer to estimate physical parameters. However, these approaches frequently necessitate extensive hyperparameter tuning for each scene and involve a costly optimization process, which limits both their practicality and generalizability. In this work, we propose a novel framework, Vid2Sim, a generalizable video-based approach for recovering geometry and physical properties through a mesh-free reduced simulation based on Linear Blend Skinning (LBS), offering high computational efficiency and versatile representation capability. Specifically, Vid2Sim first reconstructs the observed configuration of the physical system from video using a feed-forward neural network trained to capture physical world knowledge. A lightweight optimization pipeline then refines the estimated appearance, geometry, and physical properties to closely align with video observations within just a few minutes. Additionally, after the reconstruction, Vid2Sim enables high-quality, mesh-free simulation with high efficiency. Extensive experiments demo

nstrate that our method achieves superior accuracy and efficiency in reconstructing geometry and physical properties from video data.
****

SceneTAP: Scene-Coherent Typographic Adversarial Planner against Vision-Language Models in Real-World Environments
Yue Cao, Yun Xing, Jie Zhang, Di Lin, Tianwei Zhang, Ivor Tsang, Yang Liu, Qing Guo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25050-25059
Large vision-language models (LVLMs) have shown remarkable capabilities in interpreting visual content. While existing works demonstrate these models' vulnerability to deliberately placed adversarial texts, such texts are often easily identifiable as anomalous. In this paper, we present the first approach to generate scene-coherent typographic adversarial attacks that mislead advanced LVLMs while maintaining visual naturalness through the capability of the LLM-based agent. Our approach addresses three critical questions: what adversarial text to generate, where to place it within the scene, and how to integrate it seamlessly. We propose a training-free, multi-modal LLM-driven scene-coherent typographic adversarial planning (SceneTAP) that employs a three-stage process: scene understanding, adversarial planning, and seamless integration. The SceneTAP utilizes chain-of-thought reasoning to comprehend the scene, formulate effective adversarial text, strategically plan its placement, and provide detailed instructions for natural integration within the image. This is followed by a scene-coherent TextDiffuser that executes the attack using a local diffusion mechanism. We extend our method to real-world scenarios by printing and placing generated patches in physical environments, demonstrating its practical implications. Extensive experiments show that our scene-coherent adversarial text successfully misleads state-of-the-art LVLMs, including ChatGPT-4o, even after capturing new images of physical setups. Our evaluations demonstrate a significant increase in attack success rates while maintaining visual naturalness and contextual appropriateness. This work highlights vulnerabilities in current vision-language models to sophisticated, scene-coherent adversarial attacks and provides insights into potential defense mechanisms. We release our code at https://github.com/tsingqguo/scenetap.
****

Collaborative Decoding Makes Visual Auto-Regressive Modeling Efficient
Zigeng Chen, Xinyin Ma, Gongfan Fang, Xinchao Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23334-23344
In the rapidly advancing field of image generation, *Visual Auto-Regressive* (VAR) modeling has garnered considerable attention for its innovative next-scale prediction approach. This paradigm offers substantial improvements in efficiency, scalability, and zero-shot generalization. Yet, the inherently coarse-to-fine nature of VAR introduces a prolonged token sequence, leading to prohibitive memory consumption and computational redundancies. To overcome these bottlenecks, we propose *Collaborative Decoding* (CoDe), a novel decoding strategy tailored to the VAR framework. CoDe capitalizes on two critical observations: the substantially reduced parameter demands at larger scales and the exclusive generation patterns across different scales. Based on these insights, we partition the multi-scale inference process into a seamless collaboration between a large model and a small model. The large model serves as the 'drafter', specializing in generating low-frequency content at smaller scales, while the smaller model serves as the 'refiner', solely focusing on predicting high-frequency details at larger scales. This collaboration yields remarkable efficiency with minimal impact on quality: CoDe achieves a 1.7x speedup, slashes memory usage by around 50%, and preserves image quality with only a negligible FID increase from 1.95 to 1.98. When drafting steps are further decreased, CoDe can achieve an impressive 2.9x acceleration, reaching over 41 images/s at 256x256 resolution on a single NVIDIA 4090 GPU, while preserving a commendable FID of 2.27.
****

AerialMegaDepth: Learning Aerial-Ground Reconstruction and View Synthesis
Khiem Vuong, Anurag Ghosh, Deva Ramanan, Srinivasa Narasimhan, Shubham Tulsiani; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2

025, pp. 21674-21684
We explore the task of geometric reconstruction of images captured from a mixture of ground and aerial views. Current state-of-the-art learning-based approaches fail to handle the extreme viewpoint variation between aerial-ground image pairs. Our hypothesis is that the lack of high-quality, co-registered aerial-ground datasets for training is a key reason for this failure. Such data is difficult to assemble precisely because it is difficult to reconstruct in a scalable way. To overcome this challenge, we propose a scalable framework combining pseudo-synthetic renderings from 3D city-wide meshes (e.g., Google Earth) with real, ground-level crowd-sourced images (e.g., MegaDepth). The pseudo-synthetic data simulates a wide range of aerial viewpoints, while the real, crowd-sourced images help improve visual fidelity for ground-level images where mesh-based renderings lack sufficient detail, effectively bridging the domain gap between real images and pseudo-synthetic renderings. Using this hybrid dataset, we fine-tune several state-of-the-art algorithms and achieve significant improvements on real-world, zero-shot aerial-ground tasks. For example, we observe that baseline DUSt3R localizes fewer than 5% of aerial-ground pairs within 5 degrees of camera rotation error, while fine-tuning with our data raises accuracy to nearly 56%, addressing a major failure point in handling large viewpoint changes. Beyond camera estimation and scene reconstruction, our dataset also improves performance on downstream tasks like novel-view synthesis in challenging aerial-ground scenarios, demonstrating the practical value of our approach in real-world applications.
****

Towards Training-free Anomaly Detection with Vision and Language Foundation Models
Jinjin Zhang, Guodong Wang, Yizhou Jin, Di Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15204-15213
Anomaly detection is valuable for real-world applications, such as industrial quality inspection. However, most approaches focus on detecting local structural anomalies while neglecting compositional anomalies incorporating logical constraints. In this paper, we introduce LogSAD, a novel multi-modal framework that requires no training for both Logical and Structural Anomaly Detection. First, we propose a match-of-thought architecture that employs advanced large multi-modal models (i.e. GPT-4V) to generate matching proposals, formulating interests and compositional rules of thought for anomaly detection. Second, we elaborate on multi-granularity anomaly detection, consisting of patch tokens, sets of interests, and composition matching with vision and language foundation models. Subsequently, we present a calibration module to align anomaly scores from different detectors, followed by integration strategies for the final decision. Consequently, our approach addresses both logical and structural anomaly detection within a unified framework and achieves state-of-the-art results without the need for training, even when compared to supervised approaches, highlighting its robustness and effectiveness. Code is available at https://github.com/zhang0jhon/LogSAD.
****

LiVOS: Light Video Object Segmentation with Gated Linear Matching
Qin Liu, Jianfeng Wang, Zhengyuan Yang, Linjie Li, Kevin Lin, Marc Niethammer, Lijuan Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8668-8678
Semi-supervised video object segmentation (VOS) has been largely driven by space-time memory (STM) networks, which store past frame features in a spatiotemporal memory to segment the current frame via softmax attention. However, STM networks face memory limitations due to the quadratic complexity of softmax matching, restricting their applicability as video length and resolution increase. To address this, we propose LiVOS, a lightweight memory network that employs linear matching via linear attention, reformulating memory matching into a recurrent process that reduces the quadratic attention matrix to a constant-size, spatiotemporal-agnostic 2D state. To enhance selectivity, we introduce gated linear matching, where a data-dependent gate matrix is multiplied with the state matrix to control what information to retain or discard. Experiments on diverse benchmarks demonstrated the effectiveness of our method. It achieved 64.8 J&F on MOSE and 85.1 J

&F on DAVIS, surpassing all non-STM methods and narrowing the gap with STM-based approaches. For longer and higher-resolution videos, it matched STM-based methods with 53% less GPU memory and supports 4096p inference on a 32G consumer-grade GPU--a previously cost-prohibitive capability--opening the door for long and high-resolution video foundation models.
****
Dynamic Content Prediction with Motion-aware Priors for Blind Face Video Restoration
Lianxin Xie, Bingbing Zheng, Si Wu, Hau San Wong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17821-17830
Blind Face Video Restoration (BFVR) focuses on reconstructing high-quality facial image sequences from degraded video inputs. The main challenge is address unknown degradations, while maintaining temporal consistency across frames. Current blind face restoration methods are primarily designed for images, and directly applying these approaches to BFVR will encounter a significant drop in restoration performance. In this work, we proposed Dynamic Content Prediction with Motion-aware Priors, referred to as DCP-MP. We develop a motion-aware semantic dictionary by encoding the semantic information of high-quality videos into discrete elements, and capturing the motion information in terms of element relationships, which are derived from the dynamic temporal changes within videos. For the purpose of utilizing dictionary to represent the degraded video, we train a temporal-aware element predictor, conditioned on degraded content, to learn the prediction of discrete elements in dictionary. The predicted elements will be refined, conditioned on motion information captured by the motion-aware semantic dictionary, to enhance temporal coherence. To alleviate deviation from the original structure information, we propose a conditional structure feature correction module that corrects the features flowing from the encoder to the generator. Through extensive experiments, we validate the effectiveness of our design components and demonstrate the superior performance of DCP-MP in synthesizing high-quality video.
****
Polarized Color Screen Matting
Kenji Enomoto, Scott Cohen, Brian Price, TJ Rhodes; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 391-399
This paper considers the long-standing problem of extracting alpha mattes from video using a known background. While various color-based or polarization-based approaches have been studied in past decades, the problem remains ill-posed because the solutions solely rely on either color or polarization. We introduce Polarized Color Screen Matting, a single-shot, per-pixel matting theory for alpha matte and foreground color recovery using both color and polarization cues. Through a theoretical analysis of our diffuse-specular polarimetric compositing equation, we derive practical closed-form matting methods with their solvability conditions. Our theory concludes that an alpha matte can be extracted without manual corrections using off-the-shelf equipment such as an LCD monitor, polarization camera, and unpolarized lights with calibrated color. Experiments on synthetic and real-world datasets verify the validity of our theory and show the capability of our matting methods on real videos with quantitative and qualitative comparisons to color-based and polarization-based matting methods.
****
Visual Representation Learning through Causal Intervention for Controllable Image Editing
Shanshan Huang, Haoxuan Li, Chunyuan Zheng, Lei Wang, Guorui Liao, Zhili Gong, Huayi Yang, Li Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23484-23493
A key challenge for controllable image editing is that visual attributes with semantic meanings are not always independent, resulting in spurious correlations in model training. However, most existing methods ignore such issues, leading to biased causal visual representation learning and unintended changes to unrelated regions or attributes in the edited images. To bridge this gap, we propose a diffusion-based causal visual representation learning framework called CIDiffuser to capture causal representations of visual attributes based on structural causa

l models to address the spurious correlation. Specifically, we first decompose t
he image representation into a high-level semantic representation for core attri
butes of the image and a low-level stochastic representation for other random or
 less structured aspects, with the former extracted by a semantic encoder and th
e latter derived via a stochastic encoder. We then introduce a causal effect lea
rning module to capture the direct causal effect, that is, the difference of pot
ential outcomes before and after intervening on the visual attributes. In addit
ion, a diffusion-based learning strategy is designed to optimize the representat
ion learning process. Empirical evaluations on two benchmark datasets demonstrat
e that our approach significantly outperforms state-of-the-art methods, enabling
 highly controllable image editing by modifying learned visual representations.
****

Exploring the Deep Fusion of Large Language Models and Diffusion Transformers fo
r Text-to-Image Synthesis

Bingda Tang, Boyang Zheng, Sayak Paul, Saining Xie; Proceedings of the Computer
Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28586-28595

This paper does not describe a new method; instead, it provides a thorough explo
ration of an important yet understudied design space related to recent advances
in text-to-image synthesis---specifically, the deep fusion of large language mod
els (LLMs) with diffusion transformers (DiTs) for multimodal generation. Previou
s studies mainly focused on overall system performance rather than detailed comp
arisons with alternative methods, and key design details and training recipes we
re often left undisclosed. These gaps create uncertainty about the real potentia
l of this approach. To fill these gaps, we conduct an empirical study on text-to
-image generation, performing controlled comparisons with established baselines,
 analyzing important design choices, and providing a clear, reproducible recipe
for training at scale. We hope this work offers meaningful data points and pract
ical guidelines for future research in multimodal generation.
****

A Comprehensive Study of Decoder-Only LLMs for Text-to-Image Generation

Andrew Z. Wang, Songwei Ge, Tero Karras, Ming-Yu Liu, Yogesh Balaji; Proceedings
 of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 285
75-28585

Both text-to-image generation and large language models (LLMs) have made signifi
cant advancements. However, many text-to-image models still employ the somewhat
outdated T5 and CLIP as their text encoders. In this work, we investigate the ef
fectiveness of using modern decoder-only LLMs as text encoders for text-to-image
 diffusion models. We build a standardized training and evaluation pipeline that
 allows us to isolate and evaluate the effect of different text embeddings. We t
rain a total of 27 text-to-image models with 12 different text encoders to analy
ze the critical aspects of LLMs that could impact text-to-image generation, incl
uding the approaches to extract embeddings, different LLMs variants, and model s
izes. Our experiments reveal that the de facto way of using last-layer embedding
s as conditioning leads to inferior performance. Instead, we explore embeddings
from various layers and find that using layer-normalized averaging across all la
yers significantly improves alignment with complex prompts. Most LLMs with this
conditioning outperform the baseline T5 model, showing enhanced performance in a
dvanced visio-linguistic reasoning skills.
****

Exploring Sparse MoE in GANs for Text-conditioned Image Synthesis

Jiapeng Zhu, Ceyuan Yang, Kecheng Zheng, Yinghao Xu, Zifan Shi, Yifei Zhang, Qif
eng Chen, Yujun Shen; Proceedings of the Computer Vision and Pattern Recognition
 Conference (CVPR), 2025, pp. 18411-18423

Due to the difficulty in scaling up, generative adversarial networks (GANs) seem
 to be falling out of grace with the task of text-conditioned image synthesis. S
parsely activated mixture-of-experts (MoE) has recently been demonstrated as a v
alid solution to training large-scale models with limited resources. Inspired by
 this, we present Aurora, a GAN-based text-to-image generator that employs a col
lection of experts to learn feature processing, together with a sparse router to
 adaptively select the most suitable expert for each feature point. We adopt a t

wo-stage training strategy, which first learns a base model at 64x64 resolution followed by an upsampler to produce 512x512 images. Trained with only public dat a, our approach encouragingly closes the performance gap between GANs and indust ry-level diffusion models, maintaining a fast inference speed. We release the co de and checkpoints \href https://github.com/zhujiapeng/Aurora here to facilita te the community for further development.
****

Deformable Radial Kernel Splatting
Yi-Hua Huang, Ming-Xian Lin, Yang-Tian Sun, Ziyi Yang, Xiaoyang Lyu, Yan-Pei Cao , Xiaojuan Qi; Proceedings of the Computer Vision and Pattern Recognition Confer ence (CVPR), 2025, pp. 21513-21523
Recently, Gaussian splatting has emerged as a robust technique for representing 3D scenes, enabling real-time rasterization and high-fidelity rendering. However , Gaussians' inherent radial symmetry and smoothness constraints limit their abi lity to represent complex shapes, often requiring thousands of primitives to app roximate detailed geometry. We introduce Deformable Radial Kernel (DRK), which e xtends Gaussian splatting into a more general and flexible framework. Through le arnable radial bases with adjustable angles and scales, DRK efficiently models d iverse shape primitives while enabling precise control over edge sharpness and b oundary curvature. iven DRK's planar nature, we further develop accurate ray-pri mitive intersection computation for depth sorting and introduce efficient kernel culling strategies for improved rasterization efficiency. Extensive experiments demonstrate that DRK outperforms existing methods in both representation effici ency and rendering quality, achieving state-of-the-art performance while dramati cally reducing primitive count.
****

GOAL: Global-local Object Alignment Learning
Hyungyu Choi, Young Kyun Jang, Chanho Eom; Proceedings of the Computer Vision an d Pattern Recognition Conference (CVPR), 2025, pp. 4070-4079
Vision-language models like CLIP have shown impressive capabilities in aligning images and text, but they often struggle with lengthy and detailed text descript ions because of their training focus on short and concise captions. We present G OAL (Global-local Object Alignment Learning), a novel fine-tuning method that en hances CLIP's ability to handle lengthy text by leveraging both global and local semantic alignments between image and lengthy text. Our approach consists of tw o key components: Local Image-Sentence Matching (LISM), which identifies corresp onding pairs between image segments and descriptive sentences, and Token Similar ity-based Learning (TSL), which efficiently propagates local element attention t hrough these matched pairs. Evaluating GOAL on three new benchmarks for image-le ngthy text retrieval, we demonstrate significant improvements over baseline CLIP fine-tuning, establishing a simple yet effective approach for adapting CLIP to detailed textual descriptions. Through extensive experiments, we show that our m ethod's focus on local semantic alignment alongside global context leads to more nuanced and representative embeddings, particularly beneficial for tasks requir ing fine-grained understanding of lengthy text descriptions.
****

Bayesian Prompt Flow Learning for Zero-Shot Anomaly Detection
Zhen Qu, Xian Tao, Xinyi Gong, ShiChen Qu, Qiyu Chen, Zhengtao Zhang, Xingang Wa ng, Guiguang Ding; Proceedings of the Computer Vision and Pattern Recognition Co nference (CVPR), 2025, pp. 30398-30408
Recently, vision-language models (e.g. CLIP) have demonstrated remarkable perfor mance in zero-shot anomaly detection (ZSAD). By leveraging auxiliary data during training, these models can directly perform cross-category anomaly detection on target datasets, such as detecting defects on industrial product surfaces or id entifying tumors in organ tissues. Existing approaches typically construct text prompts through either manual design or the optimization of learnable prompt vec tors. However, these methods face several challenges: 1) handcrafted prompts req uire extensive expert knowledge and trial-and-error; 2) single-form learnable pr ompts struggle to capture complex anomaly semantics; and 3) an unconstrained pro mpt space limits generalization to unseen categories. To address these issues, w

e propose Bayesian Prompt Flow Learning (Bayes-PFL), which models the prompt spa ce as a learnable probability distribution from a Bayesian perspective. Specific ally, a prompt flow module is designed to learn both image-specific and image-ag nostic distributions, which are jointly utilized to regularize the text prompt s pace and improve the model's generalization on unseen categories. These learned distributions are then sampled to generate diverse text prompts, effectively cov ering the prompt space. Additionally, a residual cross-model attention (RCA) mod ule is introduced to better align dynamic text embeddings with fine-grained imag e features. Extensive experiments on 15 industrial and medical datasets demonstr ate our method's superior performance. The code is available at https://github.c om/xiaozhen228/Bayes-PFL.
****

Post-pre-training for Modality Alignment in Vision-Language Foundation Models
Shin'ya Yamaguchi, Dewei Feng, Sekitoshi Kanai, Kazuki Adachi, Daiki Chijiwa; Pr oceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025 , pp. 4256-4266
Contrastive language image pre-training (CLIP) is an essential component of buil ding modern vision-language foundation models. While CLIP demonstrates remarkabl e zero-shot performance on downstream tasks, the multi-modal feature spaces stil l suffer from a modality gap, which is a gap between image and text feature clus ters and limits downstream task performance. Although existing works attempt to address the modality gap by modifying pre-training or fine-tuning, they struggle with heavy training costs with large datasets or degradations of zero-shot perf ormance. This paper presents CLIP-Refine, a post-pre-training method for CLIP mo dels at a phase between pre-training and fine-tuning. CLIP-Refine aims to align the feature space with 1 epoch training on small image-text datasets without zer o-shot performance degradations. To this end, we introduce two techniques: rando m feature alignment (RaFA) and hybrid contrastive-distillation (HyCD). RaFA alig ns the image and text features to follow a shared prior distribution by minimizi ng the distance to random reference vectors sampled from the prior. HyCD updates the model with hybrid soft labels generated by combining ground-truth image-tex t pair labels and outputs from the pre-trained CLIP model. This contributes to a chieving both maintaining the past knowledge and learning new knowledge to align features. Our extensive experiments with multiple classification and retrieval tasks show that CLIP-Refine succeeds in mitigating the modality gap and improvin g the zero-shot performance. Code is available at https://github.com/yshinya6/cl ip-refine.
****

Efficient Event-Based Object Detection: A Hybrid Neural Network with Spatial and Temporal Attention
Soikat Hasan Ahmed, Jan Finkbeiner, Emre Neftci; Proceedings of the Computer Vis ion and Pattern Recognition Conference (CVPR), 2025, pp. 13970-13979
Event cameras offer high temporal resolution and dynamic range with minimal moti on blur, making them promising for robust object detection. While Spiking Neura l Networks (SNNs) on neuromorphic hardware are often considered for energy effic ient and low latency event-based data processing, they often fall short of Artif icial Neural Networks (ANNs) in accuracy and flexibility. Here, we introduce Att ention-based Hybrid SNN-ANN backbones for event-based object detection to levera ge the strengths of both SNN and ANN architectures. A novel Attention-based SNN-ANN bridge module proposed to captures sparse spatial and temporal relations fro m the SNN layer and converts them into dense feature maps for the ANN part of th e backbone. Additionally, we present a variant that integrates DWConvLSTMs to th e ANN blocks to capture slower dynamics. This multi-timescale network combines f ast SNN processing for short timesteps with long-term dense RNN processing, effe ctively capturing both fast and slow dynamics.Experimental results demonstrate t hat our proposed method surpasses SNN-based approaches by significant margins, w ith results comparable to existing ANN and RNN-based methods. Unlike ANN-only ne tworks, the hybrid setup allow us to implement the SNN blocks on digital neuromo rphic hardware to investigate the feasibility of our approach.Extensive ablation studies and implementation on neuromorphic hardware confirm the effectiveness o

f our proposed modules and architectural choices.Our hybrid SNN-ANN architecture s pave the way for ANN-like performance at a drastically reduced parameter, late ncy, and power budget.
****

SynthLight: Portrait Relighting with Diffusion Model by Learning to Re-render Sy nthetic Faces
Sumit Chaturvedi, Mengwei Ren, Yannick Hold-Geoffroy, Jingyuan Liu, Julie Dorsey , Zhixin Shu; Proceedings of the Computer Vision and Pattern Recognition Confere nce (CVPR), 2025, pp. 369-379
We introduce SynthLight, a diffusion model for portrait relighting. We frame ima ge relighting as a re-rendering problem, where pixels are transformed in respons e to changes in environmental lighting. Using a physically-based rendering engin e, we create a dataset to simulate this lighting-conditioned transformation wit h 3D head assets under varying lighting. We propose training and inference strat egies to bridge the gap between the synthetic and real image domains: (1) multi- task training leveraging real human portraits without lighting labels; (2) an in ference time diffusion sampling scheme based on classifier-free guidance leverag ing the input portrait to better preserve details. Our method generalizes to div erse real portraits and produces realistic illumination effects, including specu lar highlights and cast shadows, while preserving identity. Our quantitative exp eriments on light stage data demonstrate results comparable to state-of-the-art relighting methods. Our qualitative results on in-the-wild images showcase rich and unprecedented illumination effects.
****

Pseudo Visible Feature Fine-Grained Fusion for Thermal Object Detection
Ting Li, Mao Ye, Tianwen Wu, Nianxin Li, Shuaifeng Li, Song Tang, Luping Ji; Pro ceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6710-6719
Thermal object detection is a critical task in various fields, such as surveilla nce and autonomous driving. Current state-of-the-art (SOTA) models always levera ge a prior Thermal-To-Visible (T2V) translation model to obtain visible spectrum information, followed by a cross-modality aggregation module to fuse informatio n from both modalities. However, this fusion approach does not fully exploit the complementary visible spectrum information beneficial for thermal detection. To address this issue, we propose a novel cross-modal fusion method called Pseudo Visible Feature Fine-Grained Fusion (PFGF). Specifically, a graph is constructed with nodes generated from multi-level thermal features and pseudo-visual latent features produced by the T2V model. Each level of features corresponds to a sub graph. An Inter-Mamba block is proposed to perform cross-modality fusion between nodes at the lowest level; while a Cascade Knowledge Integration (CKI) strategy is used to fuse low-level fused information to high-level subgraphs in a cascad e manner. After several iterations of graph node updating, each subgraph outputs an aggregated feature to the detection head respectively. Unlike previous cross -modal fusion methods, our approach explicitly models high-level relationships b etween cross-modal data, effectively fusing different granularity information. E xperimental results demonstrate that our method achieves SOTA detection performa nce. Code is available at https://github.com/liting1018/PFGF.
****

HUNet: Homotopy Unfolding Network for Image Compressive Sensing
Feiyang Shen, Hongping Gan; Proceedings of the Computer Vision and Pattern Recog nition Conference (CVPR), 2025, pp. 12799-12808
Deep Unfolding Networks (DUNs) have risen to prominence due to their interpretab ility and superior performance for image Compressive Sensing (CS). However, exis ting DUNs still face significant issues, such as the insufficient representation capability of single-scale image information during the iterative reconstructio n phase and loss of feature information, which fundamentally limit the further e nhancement of image CS performance. In this paper, we propose Homotopy Unfolding Network (HUNet) for image CS, which enables phase-by-phase reconstruction of im ages along homotopy path. Specifically, each iteration step of the traditional h omotopy algorithm is mapped to a Multi-scale Homotopy Iterative Module (MHIM), w

hich includes U-shaped stacked window-based Transformer blocks capable of efficient feature extraction. Within the MHIM, we design the deep homotopy continuation strategy to ensure the interpretability of the homotopy algorithm and facilitate feature learning. Additionally, we introduce a dual-path feature fusion module to mitigate the loss of high-dimensional feature information during the transmission between iterative phases, thereby maximizing the preservation of details in the reconstructed image. Extensive experiments indicate that HUNet achieves superior image reconstruction results compared to existing state-of-the-art methods. The source code is available at https://github.com/ICSResearch/HUNet.
****

HalLoc: Token-level Localization of Hallucinations for Vision Language Models
Eunkyu Park, Minyeong Kim, Gunhee Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29893-29903
Hallucinations pose a significant challenge to the reliability of large vision-language models, making their detection essential for ensuring accuracy in critical applications. Current detection methods often rely on computationally intensive models, leading to high latency and resource demands. Their definitive outcomes also fail to account for real-world scenarios where the line between hallucinated and truthful information is unclear. To address these issues, we propose HalLoc, a dataset designed for efficient, probabilistic hallucination detection. It features 150K token-level annotated samples, including hallucination types, across Visual Question Answering (VQA), instruction-following, and image captioning tasks. This dataset facilitates the development of models that detect hallucinations with graded confidence, enabling more informed user interactions. Additionally, we introduce a baseline model trained on HalLoc, offering low-overhead, concurrent hallucination detection during generation. The model can be seamlessly integrated into existing VLMs, improving reliability while preserving efficiency. The prospect of a robust plug-and-play hallucination detection module opens new avenues for enhancing the trustworthiness of vision-language models in real-world applications. The HalLoc dataset and code are publicly available at: https://github.com/dbsltm/cvpr25_halloc.
****

DiffPortrait360: Consistent Portrait Diffusion for 360 View Synthesis
Yuming Gu, Phong Tran, Yujian Zheng, Hongyi Xu, Heyuan Li, Adilbek Karmanov, Hao Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26263-26273
Generating high-quality 360-degree views of human heads from single-view images is essential for enabling accessible immersive telepresence applications and scalable personalized content creation.While cutting-edge methods for full head generation are limited to modeling realistic human heads, the latest diffusion-based approaches for style-omniscient head synthesis can produce only frontal views and struggle with view consistency, preventing their conversion into true 3D models for rendering from arbitrary angles.We introduce a novel approach that generates fully consistent 360-degree head views, accommodating human, stylized, and anthropomorphic forms, including accessories like glasses and hats. Our method builds on the DiffPortrait3D framework, incorporating a custom ControlNet for back-of-head detail generation and a dual appearance module to ensure global front-back consistency. By training on continuous view sequences and integrating a back reference image, our approach achieves robust, locally continuous view synthesis. Our model can be used to produce high-quality neural radiance fields (NeRFs) for real-time, free-viewpoint rendering, outperforming state-of-the-art methods in object synthesis and 360-degree head generation for very challenging input portraits.
****

SURGEON: Memory-Adaptive Fully Test-Time Adaptation via Dynamic Activation Sparsity
Ke Ma, Jiaqi Tang, Bin Guo, Fan Dang, Sicong Liu, Zhui Zhu, Lei Wu, Cheng Fang, Ying-Cong Chen, Zhiwen Yu, Yunhao Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30514-30523
Despite the growing integration of deep models into mobile terminals, the accura

cy of these models declines significantly due to various deployment interferences. Test-time adaptation (TTA) has emerged to improve the performance of deep models by adapting them to unlabeled target data online. Yet, the significant memory cost, particularly in resource-constrained terminals, impedes the effective deployment of most backward-propagation-based TTA methods. To tackle memory constraints, we introduce SURGEON, a method that substantially reduces memory cost while preserving comparable accuracy improvements during fully test-time adaptation (FTTA) without relying on specific network architectures or modifications to the original training procedure. Specifically, we propose a novel dynamic activation sparsity strategy that directly prunes activations at layer-specific dynamic ratios during adaptation, allowing for flexible control of learning ability and memory cost in a data-sensitive manner. Among this, two metrics, Gradient Importance and Layer Activation Memory, are considered to determine the layer-wise pruning ratios, reflecting accuracy contribution and memory efficiency, respectively. Experimentally, our method surpasses the baselines by not only reducing memory usage but also achieving superior accuracy, delivering SOTA performance across diverse datasets, architectures, and tasks.

****

NVILA: Efficient Frontier Visual Language Models

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Haotian Tang, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Jinyi Hu, Sifei Liu, Ranjay Krishna, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, Yao Lu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4122-4134

Visual language models (VLMs) have made significant advances in accuracy in recent years. However, their efficiency has received much less attention. This paper introduces NVILA, a family of open VLMs designed to optimize both efficiency and accuracy. Building on top of VILA, we improve its model architecture by first scaling up the spatial and temporal resolutions, and then compressing visual tokens. This "scale-then-compress" approach enables NVILA to efficiently process high-resolution images and long videos. We also conduct a systematic investigation to enhance the efficiency of NVILA throughout its entire lifecycle, from training to deployment. NVILA matches or surpasses the accuracy of many leading open and proprietary VLMs across a wide range of image and video benchmarks. At the same time, it reduces training costs by 1.9-5.1X, prefilling latency by 1.6-2.2X, and decoding latency by 1.2-2.8X.

****

SemiETS: Integrating Spatial and Content Consistencies for Semi-Supervised End-to-end Text Spotting

Dongliang Luo, Hanshen Zhu, Ziyang Zhang, Dingkang Liang, Xudong Xie, Yuliang Liu, Xiang Bai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9329-9338

Most previous scene text spotting methods rely on high-quality manual annotations to achieve promising performance. To reduce their expensive costs, we study semi-supervised text spotting (SSTS) to exploit useful information from unlabeled images. However, directly applying existing semi-supervised methods of general scenes to SSTS will face new challenges: 1) inconsistent pseudo labels between detection and recognition tasks, and 2) sub-optimal supervisions caused by inconsistency between teacher/student. Thus, we propose a new Semi-supervised framework for End-to-end Text Spotting, namely SemiETS that leverages the complementarity of text detection and recognition. Specifically, it gradually generates reliable hierarchical pseudo labels for each task, thereby reducing noisy labels. Meanwhile, it extracts important information in locations and transcriptions from bidirectional flows to improve consistency. Extensive experiments on three datasets under various settings demonstrate the effectiveness of SemiETS on arbitrary-shaped text. For example, it outperforms previous state-of-the-art SSL methods by a large margin on end-to-end spotting (+8.7%, +5.6%, and +2.6% H-mean under 0.5%, 1%, and 2% labeled data settings on Total-Text, respectively). More importantly, it still improves upon a strongly supervised text spotter trained with plenty

of labeled data by 2.0%. Compelling domain adaptation ability shows practical potential. Moreover, our method demonstrates consistent improvement on different text spotters. Moreover, our method demonstrates consistent improvement on different text spotters. Code will be available at https://github.com/DrLuo/SemiETS.
****

See Further When Clear: Curriculum Consistency Model
Yunpeng Liu, Boxiao Liu, Yi Zhang, Xingzhong Hou, Guanglu Song, Yu Liu, Haihang You; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18103-18112
Significant advances have been made in the sampling efficiency of diffusion and flow matching models, driven by Consistency Distillation (CD), which trains a student model to mimic the output of a teacher model at a later timestep. However, we found that the knowledge discrepancy between student and teacher varies significantly across different timesteps, leading to suboptimal performance in CD.To address this issue, we propose the Curriculum Consistency Model (CCM), which stabilizes and balances the knowledge discrepancy across timesteps. Specifically, we regard the distillation process at each timestep as a curriculum and introduce a metric based on the Peak Signal-to-Noise Ratio (PSNR) to quantify the knowledge discrepancy of this curriculum, then ensure that the curriculum maintains consistent knowledge discrepancy across different timesteps by having the teacher model iterate more steps when the noise intensity is low.Our method achieves competitive single-step sampling Frechet Inception Distance (FID) scores of 1.64 on CIFAR-10 and 2.18 on ImageNet 64x64.Moreover, we have extended our method to large-scale text-to-image models and confirmed that it generalizes well to both diffusion models (Stable Diffusion XL) and flow matching models (Stable Diffusion 3). The generated samples demonstrate improved image-text alignment and semantic structure since CCM enlarges the distillation step at large timesteps and reduces the accumulated error.
****

From Slow Bidirectional to Fast Autoregressive Video Diffusion Models
Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fredo Durand, Eli Shechtman, Xun Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22963-22974
Current video diffusion models achieve impressive generation quality but struggle in interactive applications due to bidirectional attention dependencies. The generation of a single frame requires the model to process the entire sequence, including the future. We address this limitation by adapting a pretrained bidirectional diffusion transformer to an autoregressive transformer that generates frames on-the-fly. To further reduce latency, we extend distribution matching distillation (DMD) to videos, distilling 50-step diffusion model into a 4-step generator. To enable stable and high-quality distillation, we introduce a student initialization scheme based on teacher's ODE trajectories, as well as an asymmetric distillation strategy that supervises a causal student model with a bidirectional teacher. This approach effectively mitigates error accumulation in autoregressive generation, allowing long-duration video synthesis despite training on short clips. Our model achieves a total score of 84.27 on the VBench-Long benchmark, surpassing all previous video generation models. It enables fast streaming generation of high-quality videos at 9.4 FPS on a single GPU thanks to KV caching. Our approach also enables streaming video-to-video translation, image-to-video, and dynamic prompting in a zero-shot manner. We release our code and pretrained models.
****

PassionSR: Post-Training Quantization with Adaptive Scale in One-Step Diffusion based Image Super-Resolution
Libo Zhu, Jianze Li, Haotong Qin, Wenbo Li, Yulun Zhang, Yong Guo, Xiaokang Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12778-12788
Diffusion-based image super-resolution (SR) models have shown superior performance at the cost of multiple denoising steps. However, even though the denoising step has been reduced to one, they require high computational costs and storage r

equirements, making it difficult for deployment on hardware devices. To address these issues, we propose a novel post-training quantization approach with adaptive scale in one-step diffusion (OSD) image SR, PassionSR. First, we simplify OSD model to two core components, UNet and Variational Autoencoder (VAE) by removing the CLIPEncoder. Secondly, we propose Learnable Boundary Quantizer (LBQ) and Learnable Equivalent Transformation (LET) to optimize the quantization process and manipulate activation distributions for better quantization. Finally, we design a Distributed Quantization Calibration (DQC) strategy that stabilizes the training of quantized parameters for rapid convergence. Comprehensive experiments demonstrate that PassionSR with 8-bit and 6-bit obtains comparable visual results with full-precision model. Moreover, our PassionSR achieves significant advantages over recent leading low-bit quantization methods for image SR. Our code will be released

****

RainyGS: Efficient Rain Synthesis with Physically-Based Gaussian Splatting
Qiyu Dai, Xingyu Ni, Qianfan Shen, Wenzheng Chen, Baoquan Chen, Mengyu Chu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16153-16162

We consider the problem of adding dynamic rain effects to in-the-wild scenes in a physically correct manner. Recent advances in scene modeling have made significant progress, with NeRF and 3DGS techniques emerging as powerful tools for reconstructing complex scenes. However, while effective for novel view synthesis, these methods typically struggle with challenging scene editing tasks, such as physics-based rain simulation. In contrast, traditional physics-based simulations can generate realistic rain effects, such as raindrops and splashes, but they often rely on skilled artists to carefully set up high-fidelity scenes. This process lacks flexibility and scalability, limiting its applicability to broader, open-world environments. In this work, we introduce RainyGS, a novel approach that leverages the strengths of both physics-based modeling and 3DGS to generate photorealistic, dynamic rain effects in open-world scenes with physical accuracy. At the core of our method is the integration of physically-based raindrop and shallow water simulation techniques within the fast 3DGS rendering framework, enabling realistic and efficient simulations of raindrop behavior, splashes, and reflections. Our method supports synthesizing rain effects at over 30 fps, offering users flexible control over rain intensity--from light drizzles to heavy downpours. We demonstrate that RainyGS performs effectively for both real-world outdoor scenes and large-scale driving scenarios, delivering more photorealistic and physically accurate rain effects compared to state-of-the-art methods.

****

Noise Diffusion for Enhancing Semantic Faithfulness in Text-to-Image Synthesis
Boming Miao, Chunxiao Li, Xiaoxiao Wang, Andi Zhang, Rui Sun, Zizhe Wang, Yao Zhu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23575-23584

Diffusion models have achieved impressive success in generating photorealistic images, but challenges remain in ensuring precise semantic alignment with input prompts. Optimizing the initial noisy latent offers a more efficient alternative to modifying model architectures or prompt engineering for improving semantic alignment. A latest approach, InitNo, refines the initial noisy latent by leveraging attention maps; however, these maps capture only limited information, and the effectiveness of InitNo is highly dependent on the initial starting point, as it tends to converge on a local optimum near this point. To this end, this paper proposes leveraging the language comprehension capabilities of large vision-language models (LVLMs) to guide the optimization of the initial noisy latent, and introduces the Noise Diffusion process, which updates the noisy latent to generate semantically faithful images while preserving distribution consistency. Furthermore, we provide a theoretical analysis of the condition under which the update improves semantic faithfulness. Experimental results demonstrate the effectiveness and adaptability of our framework, consistently enhancing semantic alignment across various diffusion models.

****

MonoInstance: Enhancing Monocular Priors via Multi-view Instance Alignment for Neural Rendering and Reconstruction
Wenyuan Zhang, Yixiao Yang, Han Huang, Liang Han, Kanle Shi, Yu-Shen Liu, Zhizhong Han; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21642-21653
Monocular depth priors have been widely adopted by neural rendering in multi-view based tasks such as 3D reconstruction and novel view synthesis. However, due to the inconsistent prediction on each view, how to more effectively leverage monocular cues in a multi-view context remains a challenge. Current methods treat the entire estimated depth map indiscriminately, and use it as ground truth supervision, while ignoring the inherent inaccuracy and cross-view inconsistency in monocular priors. To resolve these issues, we propose MonoInstance, a general approach that explores the uncertainty of monocular depths to provide enhanced geometric priors for neural rendering and reconstruction. Our key insight lies in aligning each segmented instance depths from multiple views within a common 3D space, thereby casting the uncertainty estimation of monocular depths into a density measure within noisy point clouds. For high-uncertainty areas where depth priors are unreliable, we further introduce a constraint term that encourages the projected instances to align with corresponding instance masks on nearby views. MonoInstance is a versatile strategy which can be seamlessly integrated into various multi-view neural rendering frameworks. Our experimental results demonstrate that MonoInstance significantly improves the performance in both reconstruction and novel view synthesis under various benchmarks.
****

Three-view Focal Length Recovery From Homographies
Yaqing Ding, Viktor Kocur, Zuzana Berger Haladova, Qianliang Wu, Shen Cai, Jian Yang, Zuzana Kukelova; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11505-11514
In this paper, we propose a novel approach for recovering focal lengths from three-view homographies. By examining the consistency of normal vectors between two homographies, we derive new explicit constraints between the focal lengths and homographies using an elimination technique. We demonstrate that three-view homographies provide two additional constraints, enabling the recovery of one or two focal lengths. We discuss four possible cases, including three cameras having an unknown equal focal length, three cameras having two different unknown focal lengths, three cameras where one focal length is known, and the other two cameras have equal or different unknown focal lengths. All the problems can be converted into solving polynomials in one or two unknowns, which can be efficiently solved using Sturm sequence or hidden variable technique. Evaluation using both synthetic and real data shows that the proposed solvers are both faster and more accurate than methods relying on existing two-view solvers. The code and data are available on https://github.com/kocurvik/hf.
****

NoPain: No-box Point Cloud Attack via Optimal Transport Singular Boundary
Zezeng Li, Xiaoyu Du, Na Lei, Liming Chen, Weimin Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3492-3502
Adversarial attacks exploit the vulnerability of deep models against adversarial samples. Existing point cloud attackers are tailored to specific models, iteratively optimizing perturbations based on gradients in either a white-box or black-box setting. Despite their promising attack performance, they often struggle to produce transferable adversarial samples due to overfitting the specific parameters of surrogate models. To overcome this issue, we shift our focus to the data distribution itself and introduce a novel approach named NoPain, which employs optimal transport (OT) to identify the inherent singular boundaries of the data manifold for cross-network point cloud attacks. Specifically, we first calculate the OT mapping from noise to the target feature space, then identify singular boundaries by locating non-differentiable positions. Finally, we sample along singular boundaries to generate adversarial point clouds. Once the singular boundaries are determined, NoPain can efficiently produce adversarial samples without the need of iterative updates or guidance from the surrogate classifiers. Extensi

ve experiments demonstrate that the proposed end-to-end method outperforms basel ine approaches in terms of both transferability and efficiency, while also maint aining notable advantages even against defense strategies. Code and model are av ailable at https://github.com/cognaclee/nopain.
****

RAP: Retrieval-Augmented Personalization for Multimodal Large Language Models
Haoran Hao, Jiaming Han, Changsheng Li, Yu-Feng Li, Xiangyu Yue; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14538-1 4548
The development of large language models (LLMs) has significantly enhanced the c apabilities of multimodal LLMs (MLLMs) as general assistants. However, lack of u ser-specific knowledge still restricts their application in human's daily life. In this paper, we introduce the **R**etrieval **A**ugmented **P**ersonalization (RAP) framework for MLLMs' personalization. Starting from a general MLLM, we tur n it into a personalized assistant in three steps. (a) Remember: We design a key -value database to store user-related information, *e.g.*, user's name, avatar a nd other attributes. (b) Retrieve: When the user initiates a conversation, RAP w ill retrieve relevant information from the database using a multimodal retriever . (c) Generate: The input query and retrieved concepts' information are fed into MLLMs to generate personalized, knowledge-augmented responses. Unlike previous methods, RAP allows real-time concept editing via updating the external database . To further improve generation quality and alignment with user-specific informa tion, we design a pipeline for data collection and create a specialized dataset for personalized training of MLLMs. Based on the dataset, we train a series of M LLMs as personalized multimodal assistants. By pretraining on large-scale datase t, RAP-MLLMs can generalize to infinite visual concepts without additional finet uning. Our models demonstrate outstanding flexibility and generation quality acr oss a variety of tasks, such as personalized image captioning, question answerin g and visual recognition. The code, data and models are available at https://hoa r012.github.io/RAP-Project/.
****

FADA: Fast Diffusion Avatar Synthesis with Mixed-Supervised Multi-CFG Distillati on
Tianyun Zhong, Chao Liang, Jianwen Jiang, Gaojie Lin, Jiaqi Yang, Zhou Zhao; Pro ceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3101-3110
Diffusion-based audio-driven talking avatar methods have recently gained attenti on for their high-fidelity, vivid, and expressive results. However, their slow i nference speed limits practical applications. Despite the development of various distillation techniques for diffusion models, we found that naive diffusion dis tillation methods do not yield satisfactory results. Distilled models exhibit re duced robustness with open-set input images and a decreased correlation between audio and video compared to teacher models, undermining the advantages of diffus ion models. To address this, we propose FADA (Fast Diffusion Avatar Synthesis wi th Mixed-Supervised Multi-CFG Distillation). We first designed a mixed-supervise d loss to leverage data of varying quality and enhance the overall model capabil ity as well as robustness. Additionally, we propose a multi-CFG distillation wit h learnable tokens to utilize the correlation between audio and reference image conditions, reducing the threefold inference runs caused by multi-CFG with accep table quality degradation. Extensive experiments across multiple datasets show t hat FADA generates vivid videos comparable to recent diffusion model-based metho ds while achieving an NFE speedup of 4.17-12.5 times. Demos are available at our webpage https://fadavatar.github.io.
****

CAT4D: Create Anything in 4D with Multi-View Video Diffusion Models
Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T. Barr on, Aleksander Holynski; Proceedings of the Computer Vision and Pattern Recognit ion Conference (CVPR), 2025, pp. 26057-26068
We present CAT4D, a method for creating 4D (dynamic 3D) scenes from monocular vi deo. CAT4D leverages a multi-view video diffusion model trained on a diverse com

bination of datasets to enable novel view synthesis at any specified camera pose s and timestamps. Combined with a novel sampling approach, this model can transf orm a single monocular video into a multi-view video, enabling robust 4D reconst ruction via optimization of a deformable 3D Gaussian representation. We demonst rate competitive performance on novel view synthesis and dynamic scene reconstru ction benchmarks, and highlight the creative capabilities for 4D scene generatio n from real or generated videos. See our project page for results and interactiv e demos: cat-4d.github.io.
****

Exploring Semantic Feature Discrimination for Perceptual Image Super-Resolution and Opinion-Unaware No-Reference Image Quality Assessment
Guanglu Dong, Xiangyu Liao, Mingyang Li, Guihuan Guo, Chao Ren; Proceedings of t he Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28176-28 187
Generative Adversarial Networks (GANs) have been widely applied to image super-r esolution (SR) to enhance the perceptual quality. However, most existing GAN-bas ed SR methods typically perform coarse-grained discrimination directly on images and ignore the semantic information of images, making it challenging for the su per resolution networks (SRN) to learn fine-grained and semantic-related texture details. To alleviate this issue, we propose a semantic feature discrimination method, SFD, for perceptual SR. Specifically, we first design a feature discrimi nator (Feat-D), to discriminate the pixel-wise middle semantic features from CLI P, aligning the feature distributions of SR images with that of high-quality ima ges. Additionally, we propose a text-guided discrimination method (TG-D) by intr oducing learnable prompt pairs (LPP) in an adversarial manner to perform discrim ination on the more abstract output feature of CLIP, further enhancing the discr iminative ability of our method. With both Feat-D and TG-D, our SFD can effectiv ely distinguish between the semantic feature distributions of low-quality and hi gh-quality images, encouraging SRN to generate more realistic and semantic-relev ant textures. Furthermore, based on the trained Feat-D and LPP, we propose a nov el opinion-unaware no-reference image quality assessment (OU NR-IQA) method, SFD -IQA, greatly improving OU NR-IQA performance without any additional targeted tr aining. Extensive experiments on classical SISR, real-world SISR, and OU NR-IQA tasks, demonstrate the effectiveness of our proposed methods. Code is available at https://github.com/GuangluDong0728/SFD.
****

Distilling Long-tailed Datasets
Zhenghao Zhao, Haoxuan Wang, Yuzhang Shang, Kai Wang, Yan Yan; Proceedings of th e Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30609-306 18
Dataset distillation (DD) aims to synthesize a small information-rich dataset fr om a large dataset for efficient neural network training. However, existing data set distillation methods struggle with long-tailed datasets, which are prevalent in real-world scenarios. By investigating the reasons behind this unexpected re sult, we identified two main causes: 1) The distillation process on imbalanced d atasets develops biased gradients, leading to the synthesis of similarly imbalan ced distilled datasets. 2) The experts trained on such datasets perform suboptim ally on tail classes, resulting in misguided distillation supervision and poor-q uality soft-label initialization. To address these issues, we first propose Dist ribution-agnostic Matching to avoid directly matching the biased expert trajecto ries. It reduces the distance between the student and the biased expert trajecto ries and prevents the tail class bias from being distilled to the synthetic data set. Moreover, we improve the distillation guidance with Expert Decoupling, whic h jointly matches the decoupled backbone and classifier to improve the tail clas s performance and initialize reliable soft labels. This work pioneers the field of long-tailed dataset distillation (LTDD), marking the first effective effort t o distill long-tailed datasets. Our code will be made public at https://github.c om/ichbill/LTDD.
****

Gaze-LLE: Gaze Target Estimation via Large-Scale Learned Encoders

Fiona Ryan, Ajay Bati, Sangmin Lee, Daniel Bolya, Judy Hoffman, James M. Rehg; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28874-28884

We address the problem of gaze target estimation, which aims to predict where a person is looking in a scene. Predicting a person's gaze target requires reasoning both about the person's appearance and the contents of the scene. Prior works have developed increasingly complex, hand-crafted pipelines for gaze target estimation that carefully fuse features from separate scene encoders, head encoders, and auxiliary models for signals like depth and pose. Motivated by the success of general-purpose feature extractors on a variety of visual tasks, we propose Gaze-LLE, a novel transformer framework that streamlines gaze target estimation by leveraging features from a frozen DINOv2 encoder. We extract a single feature representation for the scene, and apply a person-specific positional prompt to decode gaze with a lightweight module. We demonstrate state-of-the-art performance across several gaze benchmarks and provide extensive analysis to validate our design choices. Our code and models are available at: http://github.com/fkryan/gazelle.

****

Distilling Spectral Graph for Object-Context Aware Open-Vocabulary Semantic Segmentation

Chanyoung Kim, Dayun Ju, Woojung Han, Ming-Hsuan Yang, Seong Jae Hwang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15033-15042

Open-Vocabulary Semantic Segmentation (OVSS) has advanced with recent vision-language models (VLMs), enabling segmentation beyond predefined categories through various learning schemes. Notably, training-free methods offer scalable, easily deployable solutions for handling unseen data, a key goal of OVSS. Yet, a critical issue persists: lack of object-level context consideration when segmenting complex objects in the challenging environment of OVSS based on arbitrary query prompts. This oversight limits models' ability to group semantically consistent elements within object and map them precisely to user-defined arbitrary classes. In this work, we introduce a novel approach that overcomes this limitation by incorporating object-level contextual knowledge within images. Specifically, our model enhances intra-object consistency by distilling spectral-driven features from vision foundation models into the attention mechanism of the visual encoder, enabling semantically coherent components to form a single object mask. Additionally, we refine the text embeddings with zero-shot object presence likelihood to ensure accurate alignment with the specific objects represented in the images. By leveraging object-level contextual knowledge, our proposed approach achieves state-of-the-art performance with strong generalizability across diverse datasets.

****

Incorporating Dense Knowledge Alignment into Unified Multimodal Representation Models

Yuhao Cui, Xinxing Zu, Wenhua Zhang, Zhongzhou Zhao, Jinyang Gao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29733-29743

Leveraging Large Language Models (LLMs) for text representation has achieved significant success, but the exploration of using Multimodal LLMs (MLLMs) for multimodal representation remains limited. Previous MLLM-based representation studies have primarily focused on unifying the embedding space while neglecting the importance of multimodal alignment. As a result, their cross-modal retrieval performance falls markedly behind that of the CLIP series models. To address this, in our work, we 1) construct DeKon5M, a contrastive learning dataset enriched with dense multimodal knowledge, which efficiently enhances multimodal alignment capabilities in representation tasks. 2) design a framework for training unified representation on MLLMs. Building upon this unified representation framework and the dense knowledge dataset DeKon5M, we developed the dense knowledge representation model DeKR on Qwen2VL. Through extensive quantitative and qualitative experiments, our results demonstrate that DeKR not only aligns text, image, video, and

text-image combinations within a unified embedding space but also achieves cross-modal retrieval performance comparable to SoTA CLIP series models. This fully validates the effectiveness of our approach and provides new insights for multimodal representation research.
****

Geometry Field Splatting with Gaussian Surfels
Kaiwen Jiang, Venkataram Sivaram, Cheng Peng, Ravi Ramamoorthi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5752-5762
Geometric reconstruction of opaque surfaces from images is a longstanding challenge in computer vision, with renewed interest from volumetric view synthesis algorithms using radiance fields. We leverage the geometry field proposed in recent work for stochastic opaque surfaces, which can then be converted to volume densities. We adapt Gaussian kernels or surfels to splat the geometry field rather than the volume, enabling precise reconstruction of opaque solids. Our first contribution is to derive an efficient and almost exact differentiable rendering algorithm for geometry fields parameterized by Gaussian surfels, while removing current approximations involving Taylor series and no self-attenuation. Next, we address the discontinuous loss landscape when surfels cluster near geometry, showing how to guarantee that the rendered color is a continuous function of the colors of the kernels, irrespective of ordering. Finally, we use latent representations with spherical harmonics encoded reflection vectors rather than spherical harmonics encoded colors to better address specular surfaces. We demonstrate significant improvement in the quality of reconstructed 3D surfaces on widely-used datasets.
****

Stereo4D: Learning How Things Move in 3D from Internet Stereo Videos
Linyi Jin, Richard Tucker, Zhengqi Li, David Fouhey, Noah Snavely, Aleksander Holynski; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10497-10509
Learning to understand dynamic 3D scenes from imagery is crucial for applications ranging from robotics to scene reconstruction. Yet, unlike other problems where large-scale supervised training has enabled rapid progress, directly supervising methods for recovering 3D motion remains challenging due to the fundamental difficulty of obtaining ground truth annotations. We present a system for mining high-quality 4D reconstructions from internet stereoscopic, wide-angle videos. Our system fuses and filters the outputs of camera pose estimation, stereo depth estimation, and temporal tracking methods into high-quality dynamic 3D reconstructions. We use this method to generate large-scale data in the form of world-consistent, pseudo-metric 3D point clouds with long-term motion trajectories. We demonstrate the utility of this data by training a variant of DUSt3r to predict structure and 3D motion from real-world image pairs, showing that training on our reconstructed data enables generalization to diverse real-world scenes. Project page and data at https://stereo4d.github.io
****

PS-EIP: Robust Photometric Stereo Based on Event Interval Profile
Kazuma Kitazawa, Takahito Aoto, Satoshi Ikehata, Tsuyoshi Takatani; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6241-6251
Recently, the energy-efficient photometric stereo method using an event camera has been proposed to recover surface normals from events triggered by changes in logarithmic Lambertian reflections under a moving directional light source. However, EventPS treats each event interval independently, making it sensitive to noise, shadows, and non-Lambertian reflections. This paper proposes Photometric Stereo based on Event Interval Profile (PS-EIP), a robust method that recovers pixelwise surface normals from a time-series profile of event intervals. By exploiting the continuity of the profile and introducing an outlier detection method based on profile shape, our approach enhances robustness against outliers from shadows and specular reflections. Experiments using real event data from 3D-printed objects demonstrate that PS-EIP significantly improves robustness to outliers c

ompared to EventPS's deep-learning variant, EventPS-FCN, without relying on deep learning.
****

GenPC: Zero-shot Point Cloud Completion via 3D Generative Priors
An Li, Zhe Zhu, Mingqiang Wei; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1308-1318
Existing point cloud completion methods, which typically depend on predefined synthetic training datasets, encounter significant challenges when applied to out-of-distribution, real-world scans. To overcome this limitation, we introduce a zero-shot completion framework, termed GenPC, designed to reconstruct high-quality real-world scans by leveraging explicit 3D generative priors.Our key insight is that recent feed-forward 3D generative models, trained on extensive internet-scale data, have demonstrated the ability to perform 3D generation from single-view images in a zero-shot setting. To harness this for completion, we first develop a Depth Prompting module that links partial point clouds with image-to-3D generative models by leveraging depth images as a stepping stone. To retain the original partial structure in the final results, we designed the Geometric Preserving Fusion module that aligns the generated shape with input by adaptively adjusting its pose and scale.Extensive experiments on widely used benchmarks validate the superiority and generalizability of our approach, bringing us a step closer to robust real-world scan completion.
****

FoundHand: Large-Scale Domain-Specific Learning for Controllable Hand Image Generation
Kefan Chen, Chaerin Min, Linguang Zhang, Shreyas Hampali, Cem Keskin, Srinath Sridhar; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17448-17460
Despite remarkable progress in image generation models, generating realistic hands remains a persistent challenge due to their complex articulation, varying viewpoints, and frequent occlusions. We present FoundHand, a large-scale domain-specific diffusion model for synthesizing single and dual hand images. To train our model, we introduce FoundHand-10M, a large-scale hand dataset with 2D keypoints and segmentation mask annotations. Our insight is to use 2D hand keypoints as a universal representation that encodes both hand articulation and camera viewpoint. FoundHand learns from image pairs to capture physically plausible hand articulations, natively enables precise control through 2D keypoints, and supports appearance control. Our model exhibits core capabilities that include the ability to repose hands, transfer hand appearance, and even synthesize novel views. This leads to zero-shot capabilities for fixing malformed hands in previously generated images, or synthesizing hand video sequences. We present extensive experiments and evaluations that demonstrate state-of-the-art performance of our method.
****

InterDyn: Controllable Interactive Dynamics with Video Diffusion Models
Rick Akkerman, Haiwen Feng, Michael J. Black, Dimitrios Tzionas, Victoria Fernández Abrevaya; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12467-12479
Predicting the dynamics of interacting objects is essential for both humans and intelligent systems. However, existing approaches are limited to simplified, toy settings and lack generalizability to complex, real-world environments. Recent advances in generative models have enabled the prediction of state transitions based on interventions, but focus on generating a single future state which neglects the continuous dynamics resulting from the interaction. To address this gap, we propose InterDyn, a novel framework that generates videos of interactive dynamics given an initial frame and a control signal encoding the motion of a driving object or actor. Our key insight is that large video generation models can act as both neural renderers and implicit physics "simulators", having learned interactive dynamics from large-scale video data. To effectively harness this capability, we introduce an interactive control mechanism that conditions the video generation process on the motion of the driving entity. Qualitative results demonstrate that InterDyn generates plausible, temporally consistent videos of comple

x object interactions while generalizing to unseen objects. Quantitative evaluat
ions show that InterDyn outperforms baselines that focus on static state transit
ions. This work highlights the potential of leveraging video generative models a
s implicit physics engines.
****

LLMDet: Learning Strong Open-Vocabulary Object Detectors under the Supervision o
f Large Language Models
Shenghao Fu, Qize Yang, Qijie Mo, Junkai Yan, Xihan Wei, Jingke Meng, Xiaohua Xi
e, Wei-Shi Zheng; Proceedings of the Computer Vision and Pattern Recognition Con
ference (CVPR), 2025, pp. 14987-14997
Recent open-vocabulary detectors achieve promising performance with abundant reg
ion-level annotated data. In this work, we show that an open-vocabulary detector
 co-training with a large language model by generating image-level detailed capt
ions for each image can further improve performance. To achieve the goal, we fir
st collect a dataset, GroundingCap-1M, wherein each image is accompanied by asso
ciated grounding labels and an image-level detailed caption. With this dataset,
we finetune an open-vocabulary detector with training objectives including a sta
ndard grounding loss and a caption generation loss. We take advantage of a large
 language model to generate both region-level short captions for each region of
interest and image-level long captions for the whole image. Under the supervisio
n of the large language model, the resulting detector, LLMDet, outperforms the b
aseline by a clear margin, enjoying superior open-vocabulary ability. Further, w
e show that the improved LLMDet can in turn build a stronger large multi-modal m
odel, achieving mutual benefits. The code, model, and dataset will be available.
****

Boost Your Human Image Generation Model via Direct Preference Optimization
Sanghyeon Na, Yonggyu Kim, Hyunjoon Lee; Proceedings of the Computer Vision and
Pattern Recognition Conference (CVPR), 2025, pp. 23551-23562
Human image generation is a key focus in image synthesis due to its broad applic
ations, but even slight inaccuracies in anatomy, pose, or details can compromise
 realism. To address these challenges, we explore Direct Preference Optimization
 (DPO), which trains models to generate preferred (winning) images while divergi
ng from non-preferred (losing) ones. However, conventional DPO methods use gener
ated images as winning images, limiting realism. To overcome this limitation, we
 propose an enhanced DPO approach that incorporates high-quality real images as
winning images, encouraging outputs to resemble real images rather than generate
d ones. However, implementing this concept is not a trivial task. Therefore, our
 approach, HG-DPO (Human image Generation through DPO), employs a novel curricul
um learning framework that gradually improves the output of the model toward gre
ater realism, making training more feasible. Furthermore, HG-DPO effectively ada
pts to personalized text-to-image tasks, generating high-quality and identity-sp
ecific images, which highlights the practical value of our approach.
****

Learning to Highlight Audio by Watching Movies
Chao Huang, Ruohan Gao, J. M. F. Tsang, Jan Kurcius, Cagdas Bilen, Chenliang Xu,
 Anurag Kumar, Sanjeel Parekh; Proceedings of the Computer Vision and Pattern Re
cognition Conference (CVPR), 2025, pp. 23925-23935
Recent years have seen a significant increase in video content creation and cons
umption. Crafting engaging content requires the careful curation of both visual
and audio elements. While visual cue curation, through techniques like optimal v
iewpoint selection or post-editing, has been central to media production, its na
tural counterpart, audio, has not undergone equivalent advancements. This often
results in a disconnect between visual and acoustic saliency. To bridge this gap
, we introduce a novel task: visually-guided acoustic highlighting, which aims t
o transform audio to deliver appropriate highlighting effects guided by the acco
mpanying video, ultimately creating a more harmonious audio-visual experience. W
e propose a flexible, transformer-based multimodal framework to solve this task.
 To train our model, we also introduce a new dataset--the muddy mix dataset, lev
eraging the meticulous audio and video crafting found in movies, which provides
a form of free supervision. We develop a pseudo-data generation process to simul

ate poorly mixed audio, mimicking real-world scenarios through a three-step process---separation, adjustment, and remixing. Our approach consistently outperforms several baselines in both quantitative and subjective evaluation. We also systematically study the impact of different types of contextual guidance and difficulty levels of the dataset.

****

Unified Uncertainty-Aware Diffusion for Multi-Agent Trajectory Modeling

Guillem Capellera, Antonio Rubio, Luis Ferraz, Antonio Agudo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22476-22486

Multi-agent trajectory modeling has primarily focused on forecasting future states, often overlooking broader tasks like trajectory completion, which are crucial for real-world applications such as correcting tracking data. Existing methods also generally predict agents' states without offering any state-wise measure of uncertainty. Moreover, popular multi-modal sampling methods lack any error probability estimates for each generated scene under the same prior observations, making it difficult to rank the predictions during inference time. We introduce U2Diff, a unified diffusion model designed to handle trajectory completion while providing state-wise uncertainty estimates jointly. This uncertainty estimation is achieved by augmenting the simple denoising loss with the negative log-likelihood of the predicted noise and propagating latent space uncertainty to the real state space. Additionally, we incorporate a Rank Neural Network in post-processing to enable error probability estimation for each generated mode, demonstrating a strong correlation with the error relative to ground truth. Our method outperforms the state-of-the-art solutions in trajectory completion and forecasting across four challenging sports datasets (NBA, Basketball-U, Football-U, Soccer-U), highlighting the effectiveness of uncertainty and error probability estimation.

****

WeGen: A Unified Model for Interactive Multimodal Generation as We Chat

Zhipeng Huang, Shaobin Zhuang, Canmiao Fu, Binxin Yang, Ying Zhang, Chong Sun, Zhizheng Zhang, Yali Wang, Chen Li, Zheng-Jun Zha; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23679-23689

Existing multimodal generative models fall short as qualified design copilots, as they often struggle to generate imaginative outputs once instructions are less detailed or lack the ability to maintain consistency with the provided references. In this work, we introduce WeGen, a model that unifies multimodal generation and understanding, and promotes their interplay in iterative generation. It can generate diverse results with high creativity for less detailed instructions. And it can progressively refine prior generation results or integrating specific contents from references following the instructions in its chat with users. During this process, it is capable of preserving consistency in the parts that the user is already satisfied with. To this end, we curate a large-scale dataset, extracted from Internet videos, containing rich object dynamics and auto-labeled dynamics descriptions by advanced foundation models to date. These two information are interleaved into a single sequence to enable WeGen to learn consistency-aware generation where the specified dynamics are generated while the consistency of unspecified content is preserved aligned with instructions. Besides, we introduce a prompt self-rewriting mechanism to enhance generation diversity. Extensive experiments demonstrate the effectiveness of unifying multimodal understanding and generation in WeGen and show it achieves state-of-the-art performance across various visual generation benchmarks. These also demonstrate the potential of WeGen as a user-friendly design copilot as desired.

****

HRAvatar: High-Quality and Relightable Gaussian Head Avatar

Dongbin Zhang, Yunfei Liu, Lijian Lin, Ye Zhu, Kangjie Chen, Minghan Qin, Yu Li, Haoqian Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26285-26296

Reconstructing animatable and high-quality 3D head avatars from monocular videos, especially with realistic relighting, is a valuable task. However, the limited

information from single-view input, combined with the complex head poses and facial movements, makes this challenging. Previous methods achieve real-time performance by combining 3D Gaussian Splatting with a parametric head model, but the resulting head quality suffers from inaccurate face tracking and limited expressiveness of the deformation model. These methods also fail to produce realistic effects under novel lighting conditions. To address these issues, we propose HRAvatar, a 3DGS-based method that reconstructs high-fidelity, relightable 3D head avatars. HRAvatar reduces tracking errors through end-to-end optimization and better captures individual facial deformations using learnable blendshapes and learnable linear blend skinning. Additionally, it decomposes head appearance into several physical properties and incorporates physically-based shading to account for environmental lighting. Extensive experiments demonstrate that HRAvatar not only reconstructs superior-quality heads but also achieves realistic visual effects under varying lighting conditions.

****

Latent Drifting in Diffusion Models for Counterfactual Medical Image Synthesis

Yousef Yeganeh, Azade Farshad, Ioannis Charisiadis, Marta Hasny, Martin Hartenberger, Björn Ommer, Nassir Navab, Ehsan Adeli; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7685-7695

Scaling by training on large datasets has been shown to enhance the quality and fidelity of image generation and manipulation with diffusion models; however, such large datasets are not always accessible in medical imaging due to cost and privacy issues, which contradicts one of the main applications of such models to produce synthetic samples where real data is scarce. Also, finetuning on pre-trained general models has been a challenge due to the distribution shift between the medical domain and the pre-trained models. Here, we propose Latent Drift (LD) for diffusion models that can be adopted for any fine-tuning method to mitigate the issues faced by the distribution shift or employed in inference time as a condition. Latent Drifting enables diffusion models to be conditioned for medical images fitted for the complex task of counterfactual image generation, which is crucial to investigate how parameters such as gender, age, and adding or removing diseases in a patient would alter the medical images. We evaluate our method on three public longitudinal benchmark datasets of brain MRI and chest X-rays for counterfactual image generation. Our results demonstrate significant performance gains in various scenarios when combined with different fine-tuning schemes. The source code of this work is available on GitHub.

****

Rethinking Spiking Self-Attention Mechanism: Implementing a-XNOR Similarity Calculation in Spiking Transformers

Yichen Xiao, Shuai Wang, Dehao Zhang, Wenjie Wei, Yimeng Shan, Xiaoli Liu, Yulin Jiang, Malu Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5444-5454

Transformers significantly raise the performance limits across various tasks, spurring research into integrating them into spiking neural networks. However, a notable performance gap remains between existing spiking Transformers and their artificial neural network counterparts. Here, we first analyze the cause of this gap and attribute it to the dot product's ineffectiveness in measuring similarity between spiking queries and keys, due to numerous non-spiking events. To address this, we propose a novel a-XNOR similarity measure tailored for spike trains. It redefines the correlation between non-spike pairs as a specific value a, effectively overcoming the limitations of dot-product similarity. Furthermore, considering the sparse nature of spike trains where spikes carry more information than non-spikes, the a-XNOR similarity correspondingly highlights the distinct importance of spikes over non-spikes. Extensive experiments demonstrate that a-XNOR similarity significantly improves performance across different spiking Transformer architectures on various static and neuromorphic datasets, further revealing the potential of spiking Transformers.

****

MagicQuill: An Intelligent Interactive Image Editing System

Zichen Liu, Yue Yu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Wen Wang, Zhiheng Li

u, Qifeng Chen, Yujun Shen; Proceedings of the Computer Vision and Pattern Recog nition Conference (CVPR), 2025, pp. 13072-13082
As a highly practical application, image editing encounters a variety of user de mands and thus prioritizes excellent ease of use. In this paper, we unveil Magic Quill, an integrated image editing system designed to support users in swiftly a ctualizing their creativity. Our system starts with a streamlined yet functional ly robust interface, enabling users to articulate their ideas (e.g., inserting e lements, erasing objects, altering color, etc.) with just a few strokes. These i nteractions are then monitored by a multimodal large language model (MLLM) to an ticipate user intentions in real time, bypassing the need for prompt entry. Fina lly, we apply the powerful diffusion prior, enhanced by a carefully learned two- branch plug-in module, to process the editing request with precise control. Plea se visit https://magic-quill.github.io to try out our system.
****

HeMoRa: Unsupervised Heuristic Consensus Sampling for Robust Point Cloud Registr ation
Shaocheng Yan, Yiming Wang, Kaiyan Zhao, Pengcheng Shi, Zhenjun Zhao, Yongjun Zh ang, Jiayuan Li; Proceedings of the Computer Vision and Pattern Recognition Conf erence (CVPR), 2025, pp. 1363-1373
Heuristic information for consensus set sampling is essential for correspondence -based point cloud registration, but existing approaches typically rely on super vised learning or expert-driven parameter tuning. In this work, we propose HeMoR a, a new unsupervised framework that trains a Heuristic information Generator (H eGen) to estimate sampling probabilities for correspondences using a Multi-order Reward Aggregator (MoRa) loss. The core of MoRa is to train HeGen through exten sive trials and feedback, enabling unsupervised learning. While this process can be implemented using policy optimization, directly applying the policy gradient to optimize HeGen presents challenges such as sensitivity to noise and low rewa rd efficiency. To address these issues, we propose a Maximal Reward Propagation (MRP) mechanism that enhances the training process by prioritizing noise-free si gnals and improving reward utilization. Experimental results show that equipped with HeMoRa, the consensus set sampler achieves improvements in both robustness and accuracy. For example, on the 3DMatch dataset with FCGF feature, the registr ation recall of our unsupervised methods (Ours+SM and Ours+SC^2) even outperform s the state-of-the-art supervised method VBreg. Our code is available at \href h ttps://github.com/Laka-3DV/HeMoRa \texttt HeMoRa .
****

Reducing Class-wise Confusion for Incremental Learning with Disentangled Manifol ds
Huitong Chen, Yu Wang, Yan Fan, Guosong Jiang, Qinghua Hu; Proceedings of the Co mputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10121-10130
Class incremental learning (CIL) aims to enable models to continuously learn new classes without catastrophically forgetting old ones. A promising direction is to learn and use prototypes of classes during incremental updates. Despite simpl icity and intuition, we find that such methods suffer from inadequate representa tion capability and unsatisfied feature overlap. These two factors cause class-w ise confusion and limited performance. In this paper, we develop a Confusion-REd uced AuTo-Encoder classifier (CREATE) for CIL. Specifically, our method employs a lightweight auto-encoder module to learn compact manifold for each class in th e latent subspace, constraining samples to be well reconstructed only on the sem antically correct auto-encoder. Thus, the representation stability and capabilit y of class distributions are enhanced, alleviating the potential class-wise conf usion problem. To further distinguish the overlapped features, we propose a conf usion-aware latent space separation loss that ensures samples are closely distri buted in their corresponding low-dimensional manifold while keeping away from th e distributions of drifted features from other classes. Our method demonstrates stronger representational capacity and discrimination ability by learning disent angled manifolds and reduces class confusion. Extensive experiments on multiple datasets and settings show that CREATE outperforms other state-of-the-art method s up to 5.41%.

****
Open-Vocabulary Functional 3D Scene Graphs for Real-World Indoor Spaces
Chenyangguang Zhang, Alexandros Delitzas, Fangjinhua Wang, Ruida Zhang, Xiangyang Ji, Marc Pollefeys, Francis Engelmann; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19401-19413
We introduce the task of predicting functional 3D scene graphs for real-world indoor environments from posed RGB-D images. Unlike traditional 3D scene graphs that focus on spatial relationships of objects, functional 3D scene graphs capture objects, interactive elements, and their functional relationships. Due to the lack of training data, we leverage foundation models, including visual language models (VLMs) and large language models (LLMs), to encode functional knowledge. We evaluate our approach on an extended SceneFun3D dataset and a newly collected dataset, FunGraph3D, both annotated with functional 3D scene graphs. Our method significantly outperforms adapted baselines, including Open3DSG and ConceptGraph, demonstrating its effectiveness in modeling complex scene functionalities. We also demonstrate downstream applications such as 3D question answering and robotic manipulation using functional 3D scene graphs. See our project page at https://openfungraph.github.io
****
Boosting Adversarial Transferability through Augmentation in Hypothesis Space
Yu Guo, Weiquan Liu, Qingshan Xu, Shijun Zheng, Shujun Huang, Yu Zang, Siqi Shen, Chenglu Wen, Cheng Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19175-19185
Adversarial examples can mislead deep neural networks with subtle perturbations, causing them to make incorrect predictions. Notably, adversarial examples crafted for one model can also deceive other models, a phenomenon known as the transferability of adversarial examples. To improve transferability, existing studies have designed increasingly complex mechanisms, but the improvements achieved remain relatively limited and are often difficult to adapt to other modalities, further restricting the scalability of these methods. In this work, we observe a mirroring relationship between model generalization and adversarial example transferability. Motivated by this observation, we propose an augmentation-based attack, called OPS (Operator-Perturbation-based Stochastic optimization), which constructs a stochastic optimization problem by input transformation operators and random perturbations, and solves this problem to generate adversarial examples with better transferability. Extensive experiments on both images and 3D point clouds demonstrate that OPS significantly outperforms existing state-of-the-art methods in terms of both performance and cost, showcasing the universality and superiority of our approach. The code is available at https://github.com/the-full/OPS.
****
AniMo: Species-Aware Model for Text-Driven Animal Motion Generation
Xuan Wang, Kai Ruan, Xing Zhang, Gaoang Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1929-1939
Text-driven motion generation has made significant strides in recent years. However, most existing works focus on human motion, largely overlooking the rich and diverse behaviors of animals. Understanding and synthesizing animal motion have important applications in wildlife conservation, animal ecology, and biomechanics. Animal motion modeling presents unique challenges due to species diversity, varied morphological structures, and different behavioral patterns in response to similar textual descriptions. To address these challenges, we propose AniMo for text-driven animal motion generation. AniMo consists of two stages: motion tokenization and text-to-motion generation. In the motion tokenization stage, we encode motions using a joint-aware spatiotemporal encoder with species-aware feature modulation, enabling the model to adapt to diverse skeletal structures across species. In the text-to-motion generation stage, we employ masked modeling to jointly learn the mappings between textual descriptions and motion tokens. Additionally, we introduce AniMo4D, a large-scale dataset containing 78,149 motion sequences and 185,435 textual descriptions across 114 animal species. Experimental results show that AniMo achieves superior performance on both the AniMo4D and An

imalML3D datasets, effectively capturing diverse morphological structures and behavioral patterns across animal species.
****

EditAR: Unified Conditional Generation with Autoregressive Models
Jiteng Mu, Nuno Vasconcelos, Xiaolong Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7899-7909
Recent progress in controllable image generation and editing is largely driven by diffusion-based methods. Although diffusion models perform exceptionally well in specific tasks with tailored designs, establishing a unified model is still challenging. In contrast, autoregressive models inherently feature a unified tokenized representation, which simplifies the creation of a single foundational model for various tasks. In this work, we propose EditAR, a single unified autoregressive framework for a variety of conditional image generation tasks, e.g., image editing, depth-to-image, edge-to-image, segmentation-to-image. The model takes both images and instructions as inputs, and predicts the edited images tokens in a vanilla next-token paradigm. To enhance the text-to-image alignment, we further propose to distill the knowledge from foundation models into the autoregressive modeling process. We evaluate its effectiveness across diverse tasks on established benchmarks, showing competitive performance to various state-of-the-art task-specific methods.
****

Instance-wise Supervision-level Optimization in Active Learning
Shinnosuke Matsuo, Riku Togashi, Ryoma Bise, Seiichi Uchida, Masahiro Nomura; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4939-4947
Active learning (AL) is a label-efficient machine learning paradigm that focuses on selectively annotating high-value instances to maximize learning efficiency. Its effectiveness can be further enhanced by incorporating weak supervision, which uses rough yet cost-effective annotations instead of exact (i.e., full) but expensive annotations. We introduce a novel AL framework, Instance-wise Supervision-Level Optimization (ISO), which not only selects the instances to annotate but also determines their optimal annotation level within a fixed annotation budget. Its optimization criterion leverages the value-to-cost ratio (VCR) of each instance while ensuring diversity among the selected instances. In classification experiments, ISO consistently outperforms traditional AL methods and surpasses a state-of-the-art AL approach that combines full and weak supervision, achieving higher accuracy at a lower overall cost.
****

ViiNeuS: Volumetric Initialization for Implicit Neural Surface Reconstruction of Urban Scenes with Limited Image Overlap
Hala Djeghim, Nathan Piasco, Moussab Bennehar, Luis Roldao, Dzmitry Tsishkou, Désiré Sidibé; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11854-11863
Neural implicit surface representation methods have recently shown impressive 3D reconstruction results. However, existing solutions struggle to reconstruct driving scenes due to their large size, highly complex nature and limited visual observation overlap.Hence, to achieve accurate reconstructions, additional supervision data such as LiDAR, strong geometric priors, and long training times are required.To tackle such limitations, we present ViiNeuS, a new hybrid implicit surface learning method that efficiently initializes the signed distance field to reconstruct large driving scenes from 2D street view images.ViiNeuS's hybrid architecture models two separate implicit fields: one representing the volumetric density of the scene, and another one representing the signed distance to the surface.To accurately reconstruct urban outdoor driving scenarios, we introduce a novel volume-rendering strategy that relies on self-supervised probabilistic density estimation to sample points near the surface and transition progressively from volumetric to surface representation. Our solution permits a proper and fast initialization of the signed distance field without relying on any geometric prior on the scene, compared to concurrent methods.By conducting extensive experiments on four outdoor driving datasets, we show that ViiNeuS can learn an accurate

and detailed 3D surface scene representation in various driving scenarios while being two times faster to train compared to previous state-of-the-art solutions.
****

Model Diagnosis and Correction via Linguistic and Implicit Attribute Editing
Xuanbai Chen, Xiang Xu, Zhihua Li, Tianchen Zhao, Pietro Perona, Qin Zhang, Yifan Xing; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14281-14292
How can we troubleshoot a deep visual model, i.e., understand why it makes certain mistakes and further take action to correct its behavior? We design a Model Diagnosis and Correction system (MDC), an automated framework that analyzes the pattern of errors, proposes candidate causes of attributes, conducts hypothesis testing via attribute editing, and ultimately generates counterfactual training samples to improve the performance of the model. Unlike previous methods, in addition to the linguistic attributes, our method also incorporates the analysis for implicit causal attributes, those cannot to be accurately described by language. To achieve this, we propose an image editing module capable of leveraging both implicit and linguistic attributes to generate counterfactual images depicting error patterns and further experimentally validate causality relationships. Lastly, we enrich the training set with synthetic samples depicting verified causal attributes and retrain the model, further boosting accuracy and robustness. Extensive experiments on fine-grained classification and face security applications demonstrate the superiority of our approach in model diagnosis and correction. Specifically, we achieve an average relative improvement of 62.01% in HTER for face security application over state-of-the-art methods.
****

BHViT: Binarized Hybrid Vision Transformer
Tian Gao, Yu Zhang, Zhiyuan Zhang, Huajun Liu, Kaijie Yin, Chengzhong Xu, Hui Kong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3563-3572
Model binarization has made significant progress in enabling real-time and energy-efficient computation for convolutional neural networks (CNN), offering a potential solution to the deployment challenges faced by Vision Transformers (ViTs) on edge devices. However, due to the structural differences between CNN and Transformer architectures, simply applying binary CNN strategies to the ViT models will lead to a significant performance drop. To tackle this challenge, we propose BHViT, a binarization-friendly hybrid ViT architecture, and its full binarization model with the guidance of three important observations. Initially, BHViT utilizes the local information interaction and hierarchical feature aggregation technique from coarse to fine levels to address redundant computations stemming from excessive tokens. Then, a novel module based on shift operations is proposed to enhance the performance of the binary Multi-Layer Perceptron (MLP) module without significantly increasing computational overhead. In addition, an innovative attention matrix binarization method based on quantization decomposition is proposed to evaluate the token's importance in the binarized attention matrix. Finally, we propose a regularization loss to address the inadequate optimization caused by the incompatibility between the weight oscillation in the binary layers and the Adam Optimizer. Extensive experimental results demonstrate that our proposed algorithm achieves SOTA performance among binary ViT methods. The source code is released at: https://github.com/IMRL/BHViT.
****

UniPhy: Learning a Unified Constitutive Model for Inverse Physics Simulation
Himangi Mittal, Peiye Zhuang, Hsin-Ying Lee, Shubham Tulsiani; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16208-16218
We propose UniPhy, a common latent-conditioned neural constitutive model that can encode the physical properties of diverse materials. At inference UniPhy allows `inverse simulation' i.e. inferring material properties by optimizing the scene-specific latent to match the available observations via differentiable simulation. In contrast to existing methods that treat such inference as system identification, UniPhy does not rely on user-specified material information. Compared t

o prior neural constitutive modeling approaches which learn instance specific networks, the shared training across materials improves both, robustness and accuracy of the estimates. We train UniPhy using simulated trajectories across diverse geometries and materials -- elastic, plasticine, sand, and fluids (Newtonian & non-Newtonian). At inference, given an object with unknown material properties, UniPhy can infer the material properties via latent optimization to match the motion observations, and can then allow re-simulating the object under diverse scenarios. We compare UniPhy against prior inverse simulation methods, and show that the inference from UniPhy enables more accurate replay and re-simulation under novel conditions.

****

STAA-SNN: Spatial-Temporal Attention Aggregator for Spiking Neural Networks
Tianqing Zhang, Kairong Yu, Xian Zhong, Hongwei Wang, Qi Xu, Qiang Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13959-13969

Spiking Neural Networks (SNNs) have gained significant attention due to their biological plausibility and energy efficiency, making them promising alternatives to Artificial Neural Networks (ANNs). However, the performance gap between SNNs and ANNs remains a substantial challenge hindering the widespread adoption of SNNs. In this paper, we propose a Spatial-Temporal Attention Aggregator SNN (STAA-SNN) framework, which dynamically focuses on and captures both spatial and temporal dependencies. First, we introduce a spike-driven self-attention mechanism specifically designed for SNNs. Additionally, we pioneeringly incorporate position encoding to integrate latent temporal relationships into the incoming features. For spatial-temporal information aggregation, we employ step attention to selectively amplify relevant features to variant steps. Finally, we implement a time-step random dropout strategy to avoid local optima. The framework demonstrates exceptional performance across diverse datasets and exhibits strong generalization capabilities. Notably, STAA-SNN achieves state-of-the-art results on neuromorphic datasets CIFAR10-DVS of 82.10% and with performances of 97.14%, 82.05% and 70.40% on the static datasets CIFAR-10, CIFAR-100 and ImageNet, respectively. Furthermore, this model exhibits improved performance ranging from 0.33% to 2.80% with fewer time steps.

****

Pathways on the Image Manifold: Image Editing via Video Generation
Noam Rotstein, Gal Yona, Daniel Silver, Roy Velich, David Bensaid, Ron Kimmel; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7857-7866

Recent advances in image editing, driven by image diffusion models, have shown remarkable progress. However, significant challenges remain, as these models often struggle to follow complex edit instructions accurately and frequently compromise fidelity by altering key elements of the original image. Simultaneously, video generation has made remarkable strides, with models that effectively function as consistent and continuous world simulators. In this paper, we propose merging these two fields by utilizing image-to-video models for image editing. We reformulate image editing as a temporal process, using pretrained video models to create smooth transitions from the original image to the desired edit. This approach traverses the image manifold continuously, ensuring consistent edits while preserving the original image's key aspects. Our approach achieves state-of-the-art results on text-based image editing, demonstrating significant improvements in both edit accuracy and image preservation. Visit our project page: https://rotsteinnoam.github.io/Frame2Frame/.

****

DeSplat: Decomposed Gaussian Splatting for Distractor-Free Rendering
Yihao Wang, Marcus Klasson, Matias Turkulainen, Shuzhe Wang, Juho Kannala, Arno Solin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 722-732

Gaussian splatting enables fast novel view synthesis in static 3D environments. However, reconstructing real-world environments remains challenging as distractors or occluders break the multi-view consistency assumption required for accurat

e 3D reconstruction. Most existing methods rely on external semantic information from pre-trained models, introducing additional computational overhead as pre-processing steps or during optimization. In this work, we propose a novel method, DeSplat, that directly separates distractors and static scene elements purely based on volume rendering of Gaussian primitives. We initialize Gaussians within each camera view for reconstructing the view-specific distractors to separately model the static 3D scene and distractors in the alpha compositing stages. DeSplat yields an explicit scene separation of static elements and distractors, achieving comparable results to prior distractor-free approaches without sacrificing rendering speed. We demonstrate DeSplat's effectiveness on three benchmark data sets for distractor-free novel view synthesis. See the project website at https://aaltoml.github.io/desplat/.

****

Knowledge Memorization and Rumination for Pre-trained Model-based Class-Incremental Learning

Zijian Gao, Wangwang Jia, Xingxing Zhang, Dulan Zhou, Kele Xu, Feng Dawei, Yong Dou, Xinjun Mao, Huaimin Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20523-20533

Class-Incremental Learning (CIL) enables models to continuously learn new classes while mitigating catastrophic forgetting. Recently, Pre-Trained Models (PTMs) have greatly enhanced CIL performance, even when fine-tuning is limited to the first task. This advantage is particularly beneficial for CIL methods that freeze the feature extractor after first-task fine-tuning, such as analytic learning-based approaches using a least squares solution-based classification head to acquire knowledge recursively. In this work, we revisit the analytical learning approach combined with PTMs and identify its limitations in adapting to new classes, leading to sub-optimal performance. To address this, we propose the Momentum-based Analytical Learning (MoAL) approach. MoAL achieves robust knowledge memorization via an analytical classification head and improves adaptivity to new classes through momentum-based adapter weight interpolation, leading to forgetting outdated knowledge. Importantly, we introduce a knowledge rumination mechanism that leverages refined adaptivity, allowing the model to revisit and reinforce old knowledge, thereby improving performance on old classes. MoAL facilitates the acquisition of new knowledge and consolidates old knowledge, achieving a win-win outcome between plasticity and stability. Extensive experiments on various incremental settings show MoAL's state-of-the-art performance

****

A Distractor-Aware Memory for Visual Object Tracking with SAM2

Jovana Videnovic, Alan Lukezic, Matej Kristan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24255-24264

Memory-based trackers are video object segmentation methods that form the target model by concatenating recently tracked frames into a memory buffer and localize the target by attending the current image to the buffered frames. While already achieving top performance on many benchmarks, it was the recent release of SAM2 that placed memory-based trackers into focus of the visual object tracking community. Nevertheless, modern trackers still struggle in the presence of distractors. We argue that a more sophisticated memory model is required, and propose a new distractor-aware memory model for SAM2 and an introspection-based update strategy that jointly addresses the segmentation accuracy as well as tracking robustness. The resulting tracker is denoted as DAM4SAM. We also propose a new distractor-distilled DiDi dataset to study the distractor problem better. DAM4SAM outperforms SAM2.1 and related SAM memory extensions on seven benchmarks and sets a solid new state-of-the-art on six of them.

****

Activating Sparse Part Concepts for 3D Class Incremental Learning

Zhenya Tian, Jun Xiao, Lupeng Liu, Haiyong Jiang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30343-30353

This work tackles the challenge of 3D Class-Incremental Learning (CIL), where a model must learn to classify new 3D objects while retaining knowledge of previously learned classes. Existing methods often struggle with catastrophic forgettin

g, misclassifying old objects due to overreliance on shortcut local features. Our approach addresses this issue by learning a set of part concepts for part-aware features. Particularly, we only activate a small subset of part concepts for the feature representation of each part-aware feature. This facilitates better generalization across categories and mitigates catastrophic forgetting. We further improve the task-wise classification through a part relation-aware Transformer design. At last, we devise learnable affinities to fuse task-wise classification heads and avoid confusion among different tasks. We evaluate our method on three 3D CIL benchmarks, achieving state-of-the-art performance. Code is available at https://github.com/zhenyatian/ILPC.

****

ProxyTransformation: Preshaping Point Cloud Manifold With Proxy Attention For 3D Visual Grounding

Qihang Peng, Henry Zheng, Gao Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24582-24592

Embodied intelligence requires agents to interact with 3D environments in real time based on language instructions. A foundational task in this domain is ego-centric 3D visual grounding. However, the point clouds rendered from RGB-D images retain a large amount of redundant background data and inherent noise, both of which can interfere with the manifold structure of the target regions. Existing point cloud enhancement methods often require a tedious process to improve the manifold, which is not suitable for real-time tasks. We propose Proxy Transformation suitable for multimodal task to efficiently improve the point cloud manifold. Our method first leverages Deformable Point Clustering to identify the point cloud sub-manifolds in target regions. Then, we propose a Proxy Attention module that utilizes multimodal proxies to guide point cloud transformation. Built upon Proxy Attention, we design a submanifold transformation generation module where textual information globally guides translation vectors for different submanifolds, optimizing relative spatial relationships of target regions. Simultaneously, image information guides linear transformations within each submanifold, refining the local point cloud manifold of target regions. Extensive experiments demonstrate that Proxy Transformation significantly outperforms all existing methods, achieving an impressive improvement of 7.49% on easy targets and 4.60% on hard targets, while reducing the computational overhead of attention blocks by 40.6%. These results establish a new SOTA in ego-centric 3D visual grounding, showcasing the effectiveness and robustness of our approach.

****

BFANet: Revisiting 3D Semantic Segmentation with Boundary Feature Analysis

Weiguang Zhao, Rui Zhang, Qiufeng Wang, Guangliang Cheng, Kaizhu Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29395-29405

3D semantic segmentation plays a fundamental and crucial role to understand 3D scenes. While contemporary state-of-the-art techniques predominantly concentrate on elevating the overall performance of 3D semantic segmentation based on general metrics (e.g. mIoU, mAcc, and oAcc), they unfortunately leave the exploration of challenging regions for segmentation mostly neglected. In this paper, we revisit 3D semantic segmentation through a more granular lens, shedding light on subtle complexities that are typically overshadowed by broader performance metrics. Concretely, we have delineated 3D semantic segmentation errors into four comprehensive categories as well as corresponding evaluation metrics tailored to each. Building upon this categorical framework, we introduce an innovative 3D semantic segmentation network called BFANet that incorporates detailed analysis of semantic boundary features. First, we design the boundary-semantic module to decouple point cloud features into semantic and boundary features, and fuse their query queue to enhance semantic features with attention. Second, we introduce a more concise and accelerated boundary pseudo-label calculation algorithm, which is 3.9 times faster than the state-of-the-art, offering compatibility with data augmentation and enabling efficient computation in training. Extensive experiments on benchmark data indicate the superiority of our BFANet model, confirming the significance of emphasizing the four uniquely designed metrics. Code is available

at https://github.com/weiguangzhao/BFANet.
****
Stable Flow: Vital Layers for Training-Free Image Editing
Omri Avrahami, Or Patashnik, Ohad Fried, Egor Nemchinov, Kfir Aberman, Dani Lischinski, Daniel Cohen-Or; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7877-7888
Diffusion models have revolutionized the field of content synthesis and editing. Recent models have replaced the traditional UNet architecture with the Diffusion Transformer (DiT), and employed flow-matching for improved training and sampling. However, they exhibit limited generation diversity. In this work, we leverage this limitation to perform consistent image edits via selective injection of attention features. The main challenge is that, unlike the UNet-based models, DiT lacks a coarse-to-fine synthesis structure, making it unclear in which layers to perform the injection. Therefore, we propose an automatic method to identify "vital layers" within DiT, crucial for image formation, and demonstrate how these layers facilitate a range of controlled stable edits, from non-rigid modifications to object addition, using the same mechanism. Next, to enable real-image editing, we introduce an improved image inversion method for flow models. Finally, we evaluate our approach through qualitative and quantitative comparisons, along with a user study, and demonstrate its effectiveness across multiple applications.
****
Video-ColBERT: Contextualized Late Interaction for Text-to-Video Retrieval
Arun Reddy, Alexander Martin, Eugene Yang, Andrew Yates, Kate Sanders, Kenton Murray, Reno Kriz, Celso M. de Melo, Benjamin Van Durme, Rama Chellappa; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19691-19701
In this work, we tackle the problem of text-to-video retrieval (T2VR). Inspired by the success of late interaction techniques in text-document, text-image, and text-video retrieval, our approach, Video-ColBERT, introduces a simple and efficient mechanism for fine-grained similarity assessment between queries and videos. Video-ColBERT is built upon 3 main components: a fine-grained spatial and temporal token-wise interaction, query and visual expansions, and a dual sigmoid loss during training. We find that this interaction and training paradigm leads to strong individual, yet compatible, representations for encoding video content. These representations lead to increases in performance on common text-to-video retrieval benchmarks compared to other bi-encoder methods.
****
Beyond Words: Augmenting Discriminative Richness via Diffusions in Unsupervised Prompt Learning
Hairui Ren, Fan Tang, He Zhao, Zixuan Wang, Dandan Guo, Yi Chang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25135-25144
Fine-tuning vision-language models (VLMs) with large amounts of unlabeled data has recently garnered significant interest. However, a key challenge remains the lack of high-quality pseudo-labeled data. Current pseudo-labeling strategies often struggle with mismatches between semantic and visual information, leading to sub-optimal performance of unsupervised prompt learning (UPL) methods.In this paper, we introduce a simple yet effective approach called Augmenting Discriminative Richness via Diffusions (AiR), toward learning a richer discriminating way to represent the class comprehensively and thus facilitate classification.Specifically, our approach includes a pseudo-label generation module that leverages high-fidelity synthetic samples to create an auxiliary classifier, which captures richer visual variation, bridging text-image-pair classification to a more robust image-image-pair classification. Additionally, we exploit the diversity of diffusion-based synthetic samples to enhance prompt learning, providing greater information for semantic-visual alignment.Extensive experiments on five public benchmarks, including RESISC45 and Flowers102, and across three learning paradigms-UL, SSL, and TRZSL-demonstrate that AiR achieves substantial and consistent performance improvements over state-of-the-art unsupervised prompt learning methods.

****
Unlocking the Potential of Unlabeled Data in Semi-Supervised Domain Generalizati
on
Dongkwan Lee, Kyomin Hwang, Nojun Kwak; Proceedings of the Computer Vision and P
attern Recognition Conference (CVPR), 2025, pp. 30599-30608
We address the problem of semi-supervised domain generalization (SSDG), where th
e distributions of train and test data differ, and only a small amount of labele
d data along with a larger amount of unlabeled data are available during trainin
g. Existing SSDG methods that leverage only the unlabeled samples for which the
model's predictions are highly confident (confident-unlabeled samples), limit th
e full utilization of the available unlabeled data. To the best of our knowledge
, we are the first to explore a method for incorporating the unconfident-unlabel
ed samples that were previously disregarded in SSDG setting. To this end, we pro
pose UPCSC to utilize these unconfident-unlabeled samples in SSDG that consists
of two modules: 1) Unlabeled Proxy-based Contrastive learning (UPC) module, trea
ting unconfident-unlabeled samples as additional negative pairs and 2) Surrogate
 Class learning (SC) module, generating positive pairs for unconfident-unlabeled
 samples using their confusing class set. These modules are plug-and-play and do
 not require any domain labels, which can be easily integrated into existing app
roaches. Experiments on four widely used SSDG benchmarks demonstrate that our ap
proach consistently improves performance when attached to baselines and outperfo
rms competing plug-and-play methods. We also analyze the role of our method in S
SDG, showing that it enhances class-level discriminability and mitigates domain
gaps.
****
TokenMotion: Decoupled Motion Control via Token Disentanglement for Human-centri
c Video Generation
Ruineng Li, Daitao Xing, Huiming Sun, Yuanzhou Ha, Jinglin Shen, Chiuman Ho; Pro
ceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025,
 pp. 1951-1961
Human-centric motion control in video generation remains a critical challenge, p
articularly when jointly controlling camera movements and human poses in scenari
os like the iconic Grammy Glambot moment. While recent video diffusion models ha
ve made significant progress, existing approaches struggle with limited motion r
epresentations and inadequate integration of camera and human motion controls. I
n this work, we present TokenMotion, the first DiT-based video diffusion framewo
rk that enables fine-grained control over camera motion, human motion, and their
 joint interaction. We represent camera trajectories and human poses as spatio-t
emporal tokens to enable local control granularity. Our approach introduces a un
ified modeling framework utilizing a decouple-and-fuse strategy, bridged by a hu
man-aware dynamic mask that effectively handles the spatially varying nature of
combined motion signals. Through extensive experiments, we demonstrate TokenMoti
on's effectiveness across both text-to-video and image-to-video paradigms, consi
stently outperforming current state-of-the-art methods in human-centric motion c
ontrol tasks. Our work represents a significant advancement in controllable vide
o generation, with particular relevance for creative production applications ran
ging from films to short-form content.
****
CholecTrack20: A Multi-Perspective Tracking Dataset for Surgical Tools
Chinedu Innocent Nwoye, Kareem Elgohary, Anvita Srinivas, Fauzan Zaid, Joël L. L
avanchy, Nicolas Padoy; Proceedings of the Computer Vision and Pattern Recogniti
on Conference (CVPR), 2025, pp. 8942-8952
Tool tracking in surgical videos is essential for advancing computer-assisted in
terventions, such as skill assessment, safety zone estimation, and human-machine
 collaboration. However, the lack of context-rich datasets limits AI application
s in this field. Existing datasets rely on overly generic tracking formalization
s that fail to capture surgical-specific dynamics, such as tools moving out of t
he camera's view or exiting the body. This results in less clinically relevant t
rajectories and a lack of flexibility for real-world surgical applications. Meth
ods trained on these datasets often struggle with visual challenges such as smok

e, reflection, and bleeding, further exposing the limitations of current approac
hes. We introduce CholecTrack20, a specialized dataset for multi-class, multi-to
ol tracking in surgical procedures. It redefines tracking formalization with thr
ee perspectives: (1) intraoperative, (2) intracorporeal, and (3) visibility, ena
bling adaptable and clinically meaningful tool trajectories. The dataset compris
es 20 full-length surgical videos, annotated at 1 fps, yielding over 35K frames
and 65K labeled tool instances. Annotations include spatial location, category,
identity, operator, phase, and visual challenges. Benchmarking state-of-the-art
methods on CholecTrack20 reveals significant performance gaps, with current appr
oaches (<45% HOTA) failing to meet the accuracy required for clinical translatio
n. These findings motivate the need for advanced and intuitive tracking algorith
ms and establish CholecTrack20 as a foundation for developing robust AI-driven s
urgical assistance systems.
****

Visual and Semantic Prompt Collaboration for Generalized Zero-Shot Learning
Huajie Jiang, Zhengxian Li, Xiaohan Yu, Yongli Hu, Baocai Yin, Jian Yang, Yuanka
i Qi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVP
R), 2025, pp. 20275-20285
Generalized zero-shot learning aims to recognize both seen and unseen classes wi
th the help of semantic information that is shared among different classes. It i
nevitably requires consistent visual-semantic alignment. Existing approaches fin
e-tune the visual backbone by seen-class data to obtain semantic-related visual
features, which may cause overfitting on seen classes with a limited number of t
raining images. This paper proposes a novel visual and semantic prompt collabora
tion framework, which utilizes prompt tuning techniques for efficient feature ad
aptation. Specifically, we design a visual prompt to integrate the visual inform
ation for discriminative feature learning and a semantic prompt to integrate the
 semantic formation for visual-semantic alignment. To achieve effective prompt i
nformation integration, we further design a weak prompt fusion mechanism for the
 shallow layers and a strong prompt fusion mechanism for the deep layers in the
network. Through the collaboration of visual and semantic prompts, we can obtain
 discriminative semantic-related features for generalized zero-shot image recogn
ition. Extensive experiments demonstrate that our framework consistently achieve
s favorable performance in both conventional zero-shot learning and generalized
zero-shot learning benchmarks compared to other state-of-the-art methods.
****

Steering Away from Harm: An Adaptive Approach to Defending Vision Language Model
 Against Jailbreaks
Han Wang, Gang Wang, Huan Zhang; Proceedings of the Computer Vision and Pattern
Recognition Conference (CVPR), 2025, pp. 29947-29957
Vision Language Models (VLMs) can produce unintended and harmful content when ex
posed to adversarial attacks, particularly because their vision capabilities cre
ate new vulnerabilities. Existing defenses, such as input preprocessing, adversa
rial training, and response evaluation-based methods, are often impractical for
real-world deployment due to their high costs. To address this challenge, we pro
pose ASTRA, an efficient and effective defense by adaptively steering models awa
y from adversarial feature directions to resist VLM attacks. Our key procedures
involve finding transferable steering vectors representing the direction of harm
ful response and applying adaptive activation steering to remove these direction
s at inference time. To create effective steering vectors, we randomly ablate th
e visual tokens from the adversarial images and identify those most strongly ass
ociated with jailbreaks. These tokens are then used to construct steering vector
s. During inference, we perform the adaptive steering method that involves the p
rojection between the steering vectors and calibrated activation, resulting in l
ittle performance drops on benign inputs while strongly avoiding harmful outputs
 under adversarial inputs. Extensive experiments across multiple models and base
lines demonstrate our state-of-the-art performance and high efficiency in mitiga
ting jailbreak risks. Additionally, ASTRA exhibits good transferability, defendi
ng against unseen attacks (i.e., structured-based attack, perturbation-based att
ack with project gradient descent variants, and text-only attack).

****
Neural LightRig: Unlocking Accurate Object Normal and Material Estimation with Multi-Light Diffusion
Zexin He, Tengfei Wang, Xin Huang, Xingang Pan, Ziwei Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26514-26524
Recovering the geometry and materials of objects from a single image is challenging due to its under-constrained nature. In this paper, we present Neural LightRig, a novel framework that boosts intrinsic estimation by leveraging auxiliary multi-lighting conditions from 2D diffusion priors. Specifically, 1) we first leverage illumination priors from large-scale diffusion models to build our multi-light diffusion model on a synthetic relighting dataset with dedicated designs. This diffusion model generates multiple consistent images, each illuminated by point light sources in different directions. 2) By using these varied lighting images to reduce estimation uncertainty, we train a large G-buffer model with a U-Net backbone to accurately predict surface normals and materials. Extensive experiments validate that our approach significantly outperforms state-of-the-art methods, enabling accurate surface normal and PBR material estimation with vivid relighting effects. Code and dataset are available on our project page at https://projects.zxhezexin.com/neural-lightrig.
****
VidMuse: A Simple Video-to-Music Generation Framework with Long-Short-Term Modeling
Zeyue Tian, Zhaoyang Liu, Ruibin Yuan, Jiahao Pan, Qifeng Liu, Xu Tan, Qifeng Chen, Wei Xue, Yike Guo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18782-18793
In this work, we systematically study music generation conditioned solely on the video. First, we present a large-scale dataset by collecting 360K video-music pairs, including various genres such as movie trailers, advertisements, and documentaries. Furthermore, we propose VidMuse, a simple framework for generating music aligned with video inputs. VidMuse stands out by producing high-fidelity music that is both acoustically and semantically aligned with the video. By incorporating local and global visual cues, VidMuse enables the creation of coherent music tracks that consistently match the video content through Long-Short-Term modeling. Through extensive experiments, VidMuse outperforms existing models in terms of audio quality, diversity, and audio-visual alignment. The code and datasets are available at https://vidmuse.github.io/
****
Human-centered Interactive Learning via MLLMs for Text-to-Image Person Re-identification
Yang Qin, Chao Chen, Zhihang Fu, Dezhong Peng, Xi Peng, Peng Hu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14390-14399
Despite remarkable advancements in text-to-image person re-identification (TIReID) facilitated by the breakthrough of cross-modal embedding models, existing methods often struggle to distinguish challenging candidate images due to intrinsic limitations, such as network architecture and data quality. To address these issues, we propose an Interactive Cross-modal Learning framework (ICL), which leverages human-centered interaction to enhance the discriminability of text queries through external multimodal knowledge. To achieve this, we propose a plug-and-play Test-time Humane-centered Interaction (TUI) module, which performs visual question answering focused on human characteristics, facilitating multi-round interactions with a multimodal large language model (MLLM) to align query intent with latent target images. Specifically, TUI refines user queries based on the MLLM responses to reduce the gap to the best-matching images, thereby boosting ranking accuracy. Additionally, to address the limitation of low-quality training texts, we introduce a novel Reorganization Data Augmentation (RDA) strategy based on information enrichment and diversity enhancement to enhance query discriminability by enriching, decomposing, and reorganizing person descriptions. Extensive experiments on four TIReID benchmarks, i.e., CUHK-PEDES, CFG-PEDES RSTPReid, RSTPReid, and UFine6926, demonstrate that our method achieves remarkable performanc

e with substantial improvement. The code will be released publicly.
****
Conditional Balance: Improving Multi-Conditioning Trade-Offs in Image Generation
Nadav Z. Cohen, Oron Nir, Ariel Shamir; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2641-2650
Balancing content fidelity and artistic style is a pivotal challenge in image generation. While traditional style transfer methods and modern Denoising Diffusion Probabilistic Models (DDPMs) strive to achieve this balance, they often struggle to do so without sacrificing either style, content, or sometimes both. This work addresses this challenge by analyzing the ability of DDPMs to maintain content and style equilibrium. We introduce a novel method to identify sensitivities within the DDPM attention layers, identifying specific layers that correspond to different stylistic aspects. By directing conditional inputs only to these sensitive layers, our approach enables fine-grained control over style and content, significantly reducing issues arising from over-constrained inputs. Our findings demonstrate that this method enhances recent stylization techniques by better aligning style and content, ultimately improving the quality of generated visual content.
****
KeyFace: Expressive Audio-Driven Facial Animation for Long Sequences via KeyFrame Interpolation
Antoni Bigata, Micha■ Stypu■kowski, Rodrigo Mira, Stella Bounareli, Konstantinos Vougioukas, Zoe Landgraf, Nikita Drobyshev, Maciej Zieba, Stavros Petridis, Maja Pantic; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5477-5488
Current audio-driven facial animation methods achieve impressive results for short videos but suffer from error accumulation and identity drift when extended to longer durations. Existing methods attempt to mitigate this through external spatial control, increasing long-term consistency but compromising the naturalness of motion. We propose KeyFace, a novel two-stage diffusion-based framework, to address these issues. In the first stage, keyframes are generated at a low frame rate, conditioned on audio input and an identity frame, to capture essential facial expressions and movements over extended periods of time. In the second stage, an interpolation model fills in the gaps between keyframes, ensuring smooth transitions and temporal coherence. To further enhance realism, we incorporate continuous emotion representations and handle a wide range of non-speech vocalizations (NSVs), such as laughter and sighs. We also introduce two new evaluation metrics for assessing lip synchronization and NSV generation. Experimental results show that KeyFace outperforms state-of-the-art methods in generating natural, coherent facial animations over extended durations, successfully encompassing NSVs and continuous emotions.
****
Context-Enhanced Memory-Refined Transformer for Online Action Detection
Zhanzhong Pang, Fadime Sener, Angela Yao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8700-8710
Online Action Detection (OAD) detects actions in streaming videos using past observations. State-of-the-art OAD approaches model past observations and their interactions with an anticipated future. The past is encoded using short- and long-term memories to capture immediate and long-range dependencies, while anticipation compensates for missing future context.We identify a training-inference discrepancy in existing OAD methods that hinders learning effectiveness. The training uses varying lengths of short-term memory, while inference relies on a full-length short-term memory. As a remedy, we propose a Context-enhanced Memory-Refined Transformer (CMeRT). CMeRT introduces a context-enhanced encoder to improve frame representations using additional near-past context. It also features a memory-refined decoder to leverage near-future generation to enhance performance. CMeRT achieves state-of-the-art in online detection and anticipation on THUMOS'14, CrossTask, and EPIC-Kitchens-100.
****
Towards Natural Language-Based Document Image Retrieval: New Dataset and Benchma

rk
Hao Guo, Xugong Qin, Jun Jie Ou Yang, Peng Zhang, Gangyan Zeng, Yubo Li, Hailun Lin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29722-29732
Document image retrieval (DIR) aims to retrieve document images from a gallery according to a given query. Existing DIR methods are primarily based on image queries that retrieve documents within the same coarse semantic category, e.g., newspapers or receipts. However, these methods struggle to effectively retrieve document images in real-world scenarios where textual queries with fine-grained semantics are usually provided. To bridge this gap, we introduce a new Natural Language-based Document Image Retrieval (NL-DIR) benchmark with corresponding evaluation metrics. In this work, natural language descriptions serve as semantically rich queries for the DIR task. The NL-DIR dataset contains 41K authentic document images, each paired with five high-quality, fine-grained semantic queries generated and evaluated through large language models in conjunction with manual verification. We perform zero-shot and fine-tuning evaluations of existing mainstream contrastive vision-language models and OCR-free visual document understanding (VDU) models. A two-stage retrieval method is further investigated for performance improvement while achieving both time and space efficiency. We hope the proposed NL-DIR benchmark can bring new opportunities and facilitate research for the VDU community. Datasets and codes will be publicly available at huggingface.co/datasets/nianbing/NL-DIR.
****
Mitigating Ambiguities in 3D Classification with Gaussian Splatting
Ruiqi Zhang, Hao Zhu, Jingyi Zhao, Qi Zhang, Xun Cao, Zhan Ma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27275-27284
3D classification with point cloud input is a fundamental problem in 3D vision. However, due to the discrete nature and the insufficient material description of point cloud representations, there are ambiguities in distinguishing wire-like and flat surfaces, as well as transparent or reflective objects. To address these issues, we propose Gaussian Splatting (GS) point cloud-based 3D classification. We find that the scale and rotation coefficients in the GS point cloud help characterize surface types. Specifically, wire-like surfaces consist of multiple slender Gaussian ellipsoids, while flat surfaces are composed of a few flat Gaussian ellipsoids. Additionally, the opacity in the GS point cloud represents the transparency characteristics of objects. As a result, ambiguities in point cloud-based 3D classification can be mitigated utilizing GS point cloud as input. To verify the effectiveness of GS point cloud input, we construct the first real-world GS point cloud dataset in the community, which includes 20 categories with 200 objects in each category. Experiments not only validate the superiority of GS point cloud input, especially in distinguishing ambiguous objects, but also demonstrate the generalization ability across different classification methods.
****
Exposure-slot: Exposure-centric Representations Learning with Slot-in-Slot Attention for Region-aware Exposure Correction
Donggoo Jung, Daehyun Kim, Guanghui Wang, Tae Hyun Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17892-17901
Image exposure correction enhances images captured under diverse real-world conditions by addressing issues of under- and over-exposure, which can result in the loss of critical details and hinder content recognition. While significant advancements have been made, current methods often fail to achieve optimal feature learning for effective correction.To overcome these challenges, we propose Exposure-slot, a novel framework that integrates a prompt-based slot-in-slot attention mechanism to cluster exposed feature regions and learn exposure-centric features for each cluster. By extending the Slot Attention algorithm with a hierarchical structure, our approach progressively clusters features, enabling precise and region-aware correction. In particular, learnable prompts tailored to exposure characteristics of slots further enhance feature quality, adapting dynamically to varying conditions. Our method delivers superior performance on benchmark datas

ets, surpassing the current state-of-the-art with a PSNR improvement of over 1.85 dB on the SICE dataset and 0.4 dB on the LCDP dataset, thereby establishing a new benchmark for multi-exposure correction. The source code can be found at:https://github.com/kdhRick2222/Exposure-slot.

****

Data-Free Group-Wise Fully Quantized Winograd Convolution via Learnable Scales

Shuokai Pan, Gerti Tuzi, Sudarshan Sreeram, Dibakar Gope; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4091-4100

Despite the revolutionary breakthroughs of large-scale text-to-image diffusion models for complex vision and downstream tasks, their extremely high computational and storage costs limit their usability. Quantization of diffusion models has been explored in recent works to reduce compute costs and memory bandwidth usage. To further improve inference time, fast convolution algorithms such as Winograd can be used for convolution layers, which account for a significant portion of computations in diffusion models. However, the significant quality loss of fully quantized Winograd using existing coarser-grained post-training quantization methods, combined with the complexity and cost of finetuning the Winograd transformation matrices for such large models to recover quality, makes them unsuitable for large-scale foundation models. Motivated by the presence of a large range of values in them, we investigate the impact of finer-grained group-wise quantization in quantizing diffusion models. While group-wise quantization can largely handle the fully quantized Winograd convolution, it struggles to deal with the large distribution imbalance in a sizable portion of the Winograd domain computation. To reduce range differences in the Winograd domain, we propose finetuning only the scale parameters of the Winograd transform matrices without using any domain-specific training data. Because our method does not depend on any training data, the generalization performance of quantized diffusion models is safely guaranteed. For text-to-image generation task, the 8-bit fully-quantized diffusion model with Winograd provides near-lossless quality (FID and CLIP scores) in comparison to the full-precision model. This, coupled with the development of highly optimized kernels for group-wise fully quantized Winograd, improves CPU wall-clock time by 31.3% when compared to the convolution layers of a diffusion model. For image classification, our method outperforms the state-of-the-art Winograd PTQ method by 1.62% and 2.56% in top-1 ImageNet accuracy on ResNet-18 and ResNet-34, respectively, with Winograd F(6, 3).

****

EdgeDiff: Edge-aware Diffusion Network for Building Reconstruction from Point Clouds

Yujun Liu, Ruisheng Wang, Shangfeng Huang, Guorong Cai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17008-17018

Building reconstruction is a challenging problem at the intersection of computer vision, photogrammetry and computer graphics. 3D wireframe presents a compelling representation for building modeling through its compact structure. Existing wireframe reconstruction methods employing vertex detection and edge regression have achieved promising results. In this paper, we develop an Edge-aware Diffusion network, dubbed EdgeDiff. As a novel paradigm for wireframe reconstruction, the EdgeDiff generates wireframe models from noise using a conditional diffusion model. During the training process, the ground truth wireframes firstly are formulated as a set of parameterized edges and then diffused into a random noise distribution. EdgeDiff learns both the noise reversal process and the network structure simultaneously. During inference, EdgeDiff iteratively refines the generated edge distribution using the denoising diffusion implicit model, enabling flexible single- or multi-step denoising and dynamic adaptation to buildings of varying complexity. Additionally, given the unique structure of wireframes, we introduce an edge attention module to extract point-wise attention from point features, using it as auxiliary information to facilitate learning of edge cues and guide the network toward improved edge awareness. Extensive experiments on the real-world Building3D dataset demonstrate that our approach achieves state-of-the-art performance.

****

GEN3C: 3D-Informed World-Consistent Video Generation with Precise Camera Control
Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, Jun Gao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6121-6132
We present GEN3C, a generative video model with precise Camera Control and temporal 3D Consistency. Prior video models already generate realistic videos, but they tend to leverage little 3D information, leading to inconsistencies, such as objects popping in and out of existence. Camera control, if implemented at all, is imprecise, because camera parameters are mere inputs to the neural network which must then infer how the video depends on the camera. In contrast, GEN3C is guided by a 3D cache: point clouds obtained by predicting the pixel-wise depth of seed images or previously generated frames. When generating the next frames, GEN3C is conditioned on the 2D renderings of the 3D cache with the new camera trajectory provided by the user. Crucially, this means that GEN3C neither has to remember what it previously generated nor does it have to infer the image structure from the camera pose. The model, instead, can focus all its generative power on previously unobserved regions, as well as advancing the scene state to the next frame. Our results demonstrate more precise camera control than prior work, as well as state-of-the-art results in sparse-view novel view synthesis, even in challenging settings such as driving scenes and monocular dynamic video.
****

A Dataset for Semantic Segmentation in the Presence of Unknowns
Zakaria Laskar, Tomas Vojir, Matej Grcic, Iaroslav Melekhov, Shankar Gangisetty, Juho Kannala, Jiri Matas, Giorgos Tolias, C.V. Jawahar; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1439-1448
Before deployment in the real-world deep neural networks require thorough evaluation of how they handle both knowns, inputs represented in the training data, and unknowns (anomalies). This is especially important for scene understanding tasks with safety critical applications, such as in autonomous driving. Existing datasets allow evaluation of only either knowns or unknowns - but not both, which is required to establish "in the wild" suitability of deep neural network models. To bridge this gap, we propose a novel anomaly segmentation dataset, ISSU, featuring a diverse set of anomaly inputs from cluttered real-world environments. The dataset is twice larger than existing anomaly segmentation datasets, and provides a training, validation and test set for controlled in-domain evaluation. The test set consists of a static and temporal part, with the later comprised of videos. The dataset provides annotations for both closed-set (knowns) and anomalies, enabling closed-set and open-set evaluation. The dataset covers diverse conditions, such as domain and cross-sensor shift, illumination variation and allows ablation of anomaly detection methods with respect to these variations. Evaluation results of current state-of-the-art methods confirm the need for improvements especially in domain-generalization, small and large object segmentation. The benchmarking code and the dataset will be made available.
****

HierarQ: Task-Aware Hierarchical Q-Former for Enhanced Video Understanding
Shehreen Azad, Vibhav Vineet, Yogesh Singh Rawat; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8545-8556
Despite advancements in multimodal large language models (MLLMs), current approaches struggle in medium-to-long video understanding due to frame and context length limitations. As a result, these models often depend on frame sampling, which risks missing key information over time and lacks task-specific relevance. To address these challenges, we introduce **HierarQ**, a task-aware hierarchical Q-Former based framework that sequentially processes frames to bypass the need for frame sampling, while avoiding LLM's context length limitations. We introduce a lightweight two-stream language-guided feature modulator to incorporate task awareness in video understanding, with the entity stream capturing frame-level object information within a short context and the scene stream identifying their broader interactions over longer period of time. Each stream is supported by dedicated memory banks which enables our proposed **Hierar**chical **Q**uerying transformer (HierarQ) to effectively capture short and long-term context. Extensive ev

aluations on **10** video benchmarks across video understanding, question answering, and captioning tasks demonstrate HierarQ's state-of-the-art performance across most datasets, proving its robustness and efficiency for comprehensive video analysis. All code will be made available upon acceptance.
****

DeNVeR: Deformable Neural Vessel Representations for Unsupervised Video Vessel Segmentation
Chun-Hung Wu, Shih-Hong Chen, Chih-Yao Hu, Hsin-Yu Wu, Kai-Hsin Chen, Yu-You Chen, Chih-Hai Su, Chih-Kuo Lee, Yu-Lun Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15682-15692
This paper presents Deformable Neural Vessel Representations (DeNVeR), an unsupervised approach for vessel segmentation in X-ray angiography videos without annotated ground truth. DeNVeR utilizes optical flow and layer separation techniques, enhancing segmentation accuracy and adaptability through test-time training. Key contributions include a novel layer separation bootstrapping technique, a parallel vessel motion loss, and the integration of Eulerian motion fields for modeling complex vessel dynamics. A significant component of this research is the introduction of the XACV dataset, the first X-ray angiography coronary video dataset with high-quality, manually labeled segmentation ground truth. Extensive evaluations on both XACV and CADICA datasets demonstrate that DeNVeR outperforms current state-of-the-art methods in vessel segmentation accuracy and generalization capability while maintaining temporal coherency. Please see our project page at kirito878.github.io/DeNVeR.
****

DH-Set: Improving Vision-Language Alignment with Diverse and Hybrid Set-Embeddings Learning
Kun Zhang, Jingyu Li, Zhe Li, S.Kevin Zhou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24993-25003
Vision-Language (VL) alignment across image and text modalities is a challenging task due to the inherent semantic ambiguity of data with multiple possible meanings. Existing methods typically solve it by learning multiple sub-representation spaces to encode each input data as a set of embeddings, and constraining diversity between whole subspaces to capture diverse semantics for accurate VL alignment. Despite their promising outcomes, existing methods suffer two imperfections: 1) actually, specific semantics is mainly expressed by some local dimensions within the subspace. Ignoring this intrinsic property, existing diversity constraints imposed on the whole subspace may impair diverse embedding learning; 2) multiple embeddings are inevitably introduced, sacrificing computational and storage efficiency. In this paper, we propose a simple yet effective Diverse and Hybrid Set-embeddings learning framework (DH-Set), which is distinct from prior work in three aspects. DH-Set 1) devises a novel semantic importance dissecting method to focus on key local dimensions within each subspace; and thereby 2) not only imposes finer-grained diversity constraint to improve the accuracy of diverse embedding learning, 3) but also mixes key dimensions of all subspaces into the single hybrid embedding to boost inference efficiency. Extensive experiments on various benchmarks and model backbones show the superiority of DH-Set over state-of-the-art methods, achieving substantial 2.3%-14.7% rSum improvements while lowering computational and storage complexity.
****

Task-Aware Clustering for Prompting Vision-Language Models
Fusheng Hao, Fengxiang He, Fuxiang Wu, Tichao Wang, Chengqun Song, Jun Cheng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14745-14755
Prompt learning has attracted widespread attention in adapting vision-language models to downstream tasks. Existing methods largely rely on optimization strategies to ensure the task-awareness of learnable prompts. Due to the scarcity of task-specific data, overfitting is prone to occur. The resulting prompts often do not generalize well or exhibit limited task-awareness. To address this issue, we propose a novel Task-Aware Clustering (TAC) framework for prompting vision-language models, which increases the task-awareness of learnable prompts by introduc

ing task-aware pre-context. The key ingredients are as follows: (a) generating t
ask-aware pre-context based on task-aware clustering that can preserve the backb
one structure of a downstream task with only a few clustering centers, (b) enhan
cing the task-awareness of learnable prompts by enabling them to interact with t
ask-aware pre-context via the well-pretrained encoders, and (c) preventing the v
isual task-aware pre-context from interfering the interaction between patch embe
ddings by masked attention mechanism. Extensive experiments are conducted on ben
chmark datasets, covering the base-to-novel, domain generalization, and cross-da
taset transfer settings. Ablation studies validate the effectiveness of key ingr
edients. Comparative results show the superiority of our TAC over competitive co
unterparts. The code is available at https://github.com/FushengHao/TAC.
****

FSboard: Over 3 Million Characters of ASL Fingerspelling Collected via Smartphon
es

Manfred Georg, Garrett Tanzer, Esha Uboweja, Saad Hassan, Maximus Shengelia, Sam
 Sepah, Sean Forbes, Thad Starner; Proceedings of the Computer Vision and Patter
n Recognition Conference (CVPR), 2025, pp. 13897-13906

Progress in machine understanding of sign languages has been slow and hampered b
y limited data. In this paper, we present FSboard, an American Sign Language fin
gerspelling dataset situated in a mobile text entry use case, collected from 147
 paid and consenting Deaf signers using Pixel 4A selfie cameras in a variety of
environments. Fingerspelling recognition is an incomplete solution that is only
one small part of sign language translation, but it could provide some immediate
 benefit to Deaf/Hard of Hearing signers while more broadly capable technology d
evelops. At >3 million characters in length and >250 hours in duration, FSboard
is the largest fingerspelling recognition dataset to date by a factor of >10x. A
s a simple baseline, we finetune 30 Hz MediaPipe Holistic landmark inputs into B
yT5-Small and achieve 11.1% Character Error Rate (CER) on a test set with unique
 phrases and signers. This quality degrades gracefully when decreasing frame rat
e and excluding face/body landmarks---plausible optimizations to help with on-de
vice performance---but falls short of human performance measured at 2.2% CER.
****

CASP: Compression of Large Multimodal Models Based on Attention Sparsity

Mohsen Gholami, Mohammad Akbari, Kevin Cannons, Yong Zhang; Proceedings of the C
omputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9372-9381

In this work, we propose an extreme compression technique for Large Multimodal M
odels (LMMs). While previous studies have explored quantization as an efficient
post-training compression method for Large Language Models (LLMs), low-bit compr
ession for multimodal models remains under-explored. The redundant nature of inp
uts in multimodal models results in a highly sparse attention matrix. We theoret
ically and experimentally demonstrate that the attention matrix's sparsity bound
s the compression error of the Query and Key weight matrices. Based on this, we
introduce CASP, a model compression technique for LMMs. Our approach performs a
data-aware low-rank decomposition on the Query and Key weight matrix, followed b
y quantization across all layers based on an optimal bit allocation process. CAS
P is compatible with any quantization technique and enhances state-of-the-art 2-
bit quantization methods (AQLM and QuIP#) by an average of 21% on image- and vid
eo-language benchmarks. The code is provided in the supplementary materials.
****

UNIC-Adapter: Unified Image-instruction Adapter with Multi-modal Transformer for
 Image Generation

Lunhao Duan, Shanshan Zhao, Wenjun Yan, Yinglun Li, Qing-Guo Chen, Zhao Xu, Weih
ua Luo, Kaifu Zhang, Mingming Gong, Gui-Song Xia; Proceedings of the Computer Vi
sion and Pattern Recognition Conference (CVPR), 2025, pp. 7963-7973

Recently, text-to-image generation models have achieved remarkable advancements,
 particularly with diffusion models facilitating high-quality image synthesis fr
om textual descriptions. However, these models often struggle with achieving pre
cise control over pixel-level layouts, object appearances, and global styles whe
n using text prompts alone. To mitigate this issue, previous works introduce con
ditional images as auxiliary inputs for image generation, enhancing control but

typically necessitating specialized models tailored to different types of refere nce inputs. In this paper, we explore a new approach to unify controllable gener ation within a single framework. Specifically, we propose the unified image-inst ruction adapter (UNIC-Adapter) built on the Multi-Modal-Diffusion Transformer ar chitecture, to enable flexible and controllable generation across diverse condit ions without the need for multiple specialized models. Our UNIC-Adapter effectiv ely extracts multi-modal instruction information by incorporating both condition al images and task instructions, injecting this information into the image gener ation process through a cross-attention mechanism enhanced by Rotary Position Em bedding. Experimental results across a variety of tasks, including pixel-level s patial control, subject-driven image generation, and style-image-based image syn thesis, demonstrate the effectiveness of our UNIC-Adapter in unified controllabl e image generation.
****

Towards Cost-Effective Learning: A Synergy of Semi-Supervised and Active Learnin g

Tianxiang Yin, Ningzhong Liu, Han Sun; Proceedings of the Computer Vision and Pa ttern Recognition Conference (CVPR), 2025, pp. 10163-10172

Active learning (AL) and semi-supervised learning (SSL) both aim to reduce annot ation costs: AL selectively annotates high-value samples from the unlabeled data , while SSL leverages abundant unlabeled data to improve model performance. Alth ough these two appear intuitively compatible, directly combining them remains ch allenging due to fundamental differences in their frameworks. Current semi-super vised active learning (SSAL) methods often lack theoretical foundations and ofte n design AL strategies tailored to a specific SSL algorithm rather than genuinel y integrating the two fields.In this paper, we incorporate AL objectives into th e overall risk formulation within the mainstream pseudo-label-based SSL framewor k, clarifying key differences between SSAL and traditional AL scenarios. To brid ge these gaps, we propose a feature re-alignment module that aligns the features of unlabeled data under different augmentations by leveraging clustering and co nsistency constraints. Experimental results demonstrate that our module enables flexible combinations of SOTA methods from both AL and SSL, yielding more effici ent algorithm performance.
****

Unveil Inversion and Invariance in Flow Transformer for Versatile Image Editing

Pengcheng Xu, Boyuan Jiang, Xiaobin Hu, Donghao Luo, Qingdong He, Jiangning Zhan g, Chengjie Wang, Yunsheng Wu, Charles Ling, Boyu Wang; Proceedings of the Compu ter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28479-28489

Leveraging the large generative prior of the flow transformer for tuning-free im age editing requires authentic inversion to project the image into the model's d omain and a flexible invariance control mechanism to preserve non-target content s. However, the prevailing diffusion inversion performs deficiently in flow-base d models, and the invariance control cannot reconcile diverse rigid and non-rigi d editing tasks. To address these, we systematically analyze the inversion and i nvariance control based on the flow transformer. Specifically, we unveil that th e Euler inversion shares a similar structure to DDIM yet is more susceptible to the approximation error. Thus, we propose a two-stage inversion to first refine the velocity estimation and then compensate for the leftover error, which pivots closely to the model prior and benefits editing. Meanwhile, we propose the inva riance control that manipulates the text features within the adaptive layer norm alization, connecting the changes in the text prompt to image semantics. This me chanism can simultaneously preserve the non-target contents while allowing rigid and non-rigid manipulation, enabling a wide range of editing types. Experiments on various scenarios demonstrate that our framework achieves flexible and accur ate editing, unlocking the potential of the flow transformer for versatile image editing.
****

Light Transport-aware Diffusion Posterior Sampling for Single-View Reconstructio n of 3D Volumes

Ludwic Leonard, Nils Thurey, Rüdiger Westermann; Proceedings of the Computer Vis

ion and Pattern Recognition Conference (CVPR), 2025, pp. 16163-16174
We introduce a single-view reconstruction technique of volumetric fields in whic h multiple light scattering effects are omnipresent, such as in clouds. We model the unknown distribution of volumetric fields using an unconditional diffusion model trained on a novel benchmark dataset comprising 1,000 synthetically simula ted volumetric density fields. The neural diffusion model is trained on the late nt codes of a novel, diffusion-friendly, monoplanar representation. The generati ve model is used to incorporate a tailored parametric diffusion posterior sampli ng technique into different reconstruction tasks. A physically-based differentia ble volume renderer is employed to provide gradients with respect to light trans port in the latent space. This stands in contrast to classic NeRF approaches and makes the reconstructions better aligned with observed data. Through various ex periments, we demonstrate single-view reconstruction of volumetric clouds at a p reviously unattainable quality.
****

DyCON: Dynamic Uncertainty-aware Consistency and Contrastive Learning for Semi-s upervised Medical Image Segmentation
Maregu Assefa, Muzammal Naseer, Iyyakutti Iyappan Ganapathi, Syed Sadaf Ali, Moh amed L Seghier, Naoufel Werghi; Proceedings of the Computer Vision and Pattern R ecognition Conference (CVPR), 2025, pp. 30850-30860
Semi-supervised learning in medical image segmentation leverages unlabeled data to reduce annotation burdens through consistency learning. However, current meth ods struggle with class imbalance and high uncertainty from pathology variations , leading to inaccurate segmentation in 3D medical images. To address these chal lenges, we present DyCON, a Dynamic Uncertainty-aware Consistency and Contrastiv e Learning framework that enhances the generalization of consistency methods wit h two complementary losses: Uncertainty-aware Consistency Loss (UnCL) and Focal Entropy-aware Contrastive Loss (FeCL). UnCL enforces global consistency by dynam ically weighting the contribution of each voxel to the consistency loss based on its uncertainty, preserving high-uncertainty regions instead of filtering them out. Initially, UnCL prioritizes learning from uncertain voxels with lower penal ties, encouraging the model to explore challenging regions. As training progress , the penalty shift towards confident voxels to refine predictions and ensure gl obal consistency. Meanwhile, FeCL enhances local feature discrimination in imbal anced regions by introducing dual focal mechanisms and adaptive confidence adjus tments into the contrastive principle. These mechanisms jointly prioritizes hard positives and negatives while focusing on uncertain sample pairs, effectively c apturing subtle lesion variations under class imbalance. Extensive evaluations o n four diverse medical image segmentation datasets (ISLES'22, BraTS'19, LA, Panc reas) show DyCON's superior performance against SOTA methods(\href https://dycon 25.github.io/) Project Page: https://dycon25.github.io/ .
****

STiL: Semi-supervised Tabular-Image Learning for Comprehensive Task-Relevant Inf ormation Exploration in Multimodal Classification
Siyi Du, Xinzhe Luo, Declan P. O'Regan, Chen Qin; Proceedings of the Computer Vi sion and Pattern Recognition Conference (CVPR), 2025, pp. 15549-15559
Multimodal image-tabular learning is gaining attention, yet it faces challenges due to limited labeled data. While earlier work has applied self-supervised lear ning (SSL) to unlabeled data, its task-agnostic nature often results in learning suboptimal features for downstream tasks. Semi-supervised learning (SemiSL), wh ich combines labeled and unlabeled data, offers a promising solution. However, e xisting multimodal SemiSL methods typically focus on unimodal or modality-shared features, ignoring valuable task-relevant modality-specific information, leadin g to a Modality Information Gap. In this paper, we propose STiL, a novel SemiSL tabular-image framework that addresses this gap by comprehensively exploring tas k-relevant information. STiL features a new disentangled contrastive consistency module to learn cross-modal invariant representations of shared information whi le retaining modality-specific information via disentanglement. We also propose a novel consensus-guided pseudo-labeling strategy to generate reliable pseudo-la bels based on classifier consensus, along with a new prototype-guided label smoo

thing technique to refine pseudo-label quality with prototype embeddings, thereby enhancing task-relevant information learning in unlabeled data. Experiments on natural and medical image datasets show that STiL outperforms the state-of-the-art supervised/SSL/SemiSL image/multimodal approaches. Our code is available at https://github.com/siyi-wind/STiL.

****

DUNE: Distilling a Universal Encoder from Heterogeneous 2D and 3D Teachers

Mert Bülent Sar■y■ld■z, Philippe Weinzaepfel, Thomas Lucas, Pau de Jorge, Diane Larlus, Yannis Kalantidis; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30084-30094

Recent multi-teacher distillation methods have unified the encoders of multiple foundation models into a single encoder, achieving competitive performance on core vision tasks like classification, segmentation, and depth estimation. This led us to ask: Could similar success be achieved when the pool of teachers also includes vision models specialized in diverse tasks across both 2D and 3D perception? In this paper, we define and investigate the problem of heterogeneous teacher distillation, or co-distillation, a challenging multi-teacher distillation scenario where teacher models vary significantly in both (a) their design objectives and (b) the data they were trained on. We explore data-sharing strategies and teacher-specific encoding, and introduce DUNE, a single encoder excelling in 2D vision, 3D understanding, and 3D human perception. Our model achieves performance comparable to that of its larger teachers, sometimes even outperforming them, on their respective tasks. Notably, DUNE surpasses MASt3R in Map-free Visual Relocalization with a much smaller encoder.

****

Black Hole-Driven Identity Absorbing in Diffusion Models

Muhammad Shaheryar, Jong Taek Lee, Soon Ki Jung; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28544-28554

Recent advances in diffusion models have positioned them as powerful generative frameworks for high-resolution image synthesis across diverse domains. The emerging h-space within these models, defined by bottleneck activations in the denoiser, offers promising pathways for semantic image editing similar to GAN latent spaces. However, as demand grows for content erasure and concept removal, privacy concerns highlight the need for identity disentanglement in the latent space of diffusion models. The high dimensional latent space poses challenges for identity removal, as traversing with random or orthogonal directions often leads to semantically unvalidated regions, resulting in unrealistic outputs.To address these issues, we propose Black Hole Driven Identity Absorption (BIA), a novel approach for identity erasure within the latent space of diffusion models. BIA uses a black hole metaphor, where the latent region representing a specified identity acts as an attractor, drawing in nearby latent points of surrounding identities to wrap the black hole. Instead of relying on random traversals for optimization, BIA employs an identity absorption mechanism by attracting and wrapping nearby validated latent points associated with other identities to achieve a vanishing effect for specified identity. Our method effectively prevents the generation of a specified identity while preserving other attributes, as validated by improved scores on identity similarity SID, FID metrics, qualitative evaluations, and user studies as compared to SOTA.

****

HiRes-LLaVA: Restoring Fragmentation Input in High-Resolution Large Vision-Language Models

Runhui Huang, Xinpeng Ding, Chunwei Wang, Jianhua Han, Yulong Liu, Hengshuang Zhao, Hang Xu, Lu Hou, Wei Zhang, Xiaodan Liang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29814-29824

High-resolution image inputs allow Large Vision-Language Models (LVLMs) to capture finer visual details, improving comprehension. However, the increased training and computational costs associated with such inputs pose significant challenges. A common approach to mitigate these costs involves slicing the input into uniform patches using sliding windows, each aligned with the vision encoder's input size. While efficient, this method fragments the input, disrupting the continui

ty of context, which negatively impacts cross-patch perception tasks. To address these limitations, we propose HiRes-LLaVA, a novel framework designed to efficiently process high-resolution inputs of any size without altering the original contextual and geometric information. HiRes-LLaVA introduces two key components: (i) a SliceRestore Adapter (SRA) that reconstructs sliced patches into their original form, enabling efficient extraction of both global and local features through down-up-sampling and convolutional layers, and (ii) a Self-Mining Sampler (SMS) that compresses visual tokens based on internal relationships, preserving original context and positional information while reducing training overhead. To assess the ability of handling context fragmentation, we construct a new benchmark, EntityGrid-QA, consisting of edge-related tasks. Extensive experiments demonstrate the superiority of HiRes-LLaVA on both existing public benchmarks and EntityGrid-QA. For example, with SRA, our method achieves a performance improvement of 12% over state-of-the-art LVLMs in addressing fragmentation issues. Additionally, our SMS outperforms other visual token downsamplers, while offering high data efficiency.

****

Hallo3: Highly Dynamic and Realistic Portrait Image Animation with Video Diffusion Transformer

Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, Siyu Zhu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21086-21095

Existing methodologies for animating portrait images face significant challenges, particularly in handling non-frontal perspectives, rendering dynamic objects around the portrait, and generating immersive, realistic backgrounds. In this paper, we introduce the first application of a pretrained transformer-based video generative model that demonstrates strong generalization capabilities and generates highly dynamic, realistic videos for portrait animation, effectively addressing these challenges. The adoption of a new video backbone model makes previous U-Net-based methods for identity maintenance, audio conditioning, and video extrapolation inapplicable. To address this limitation, we design an identity reference network consisting of a causal 3D VAE combined with a stacked series of transformer layers, ensuring consistent facial identity across video sequences. Additionally, we investigate various speech audio conditioning and motion frame mechanisms to enable the generation of continuous video driven by speech audio. Our method is validated through experiments on benchmark and newly proposed wild data sets, demonstrating substantial improvements over prior methods in generating realistic portraits characterized by diverse orientations within dynamic and immersive scenes.

****

Generative Photography: Scene-Consistent Camera Control for Realistic Text-to-Image Synthesis

Yu Yuan, Xijun Wang, Yichen Sheng, Prateek Chennuri, Xingguang Zhang, Stanley Chan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7920-7930

Image generation today can produce somewhat realistic images from text prompts. However, if one asks the generator to synthesize a specific camera setting such as creating different fields of view using a 24mm lens versus a 70mm lens, the generator will not be able to interpret and generate scene-consistent images. This limitation not only hinders the adoption of generative tools in professional photography but also highlights the broader challenge of aligning data-driven models with real-world physical settings. In this paper, we introduce Generative Photography, a framework that allows controlling camera intrinsic settings during content generation. The core innovation of this work are the concepts of Dimensionality Lifting and Differential Camera Intrinsics Learning, enabling smooth and consistent transitions across different camera settings. Experimental results show that our method produces significantly more scene-consistent photorealistic images than state-of-the-art models such as Stable Diffusion 3 and FLUX. Our code and additional results are available at https://generative-photography.github.io/project.

****
Advancing Manga Analysis: Comprehensive Segmentation Annotations for the Manga109 Dataset
Minshan Xie, Jian Lin, Hanyuan Liu, Chengze Li, Tien-Tsin Wong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8869-8878
Manga, a popular form of multimodal artwork, has traditionally been overlooked in deep learning advancements due to the absence of a robust dataset and comprehensive annotation. Manga segmentation is the key to the digital migration of manga. There exists a significant domain gap between the manga and the natural images, that fails most existing learning-based methods. To address this gap, we introduce an augmented segmentation annotation for the Manga109 dataset, a collection of 109 manga volumes, that offers intricate artworks in a rich variety of styles. We introduce a detailed annotation that extends beyond the original simple bounding boxes to the segmentation masks with pixel-level precision. It provides object category, location, and instance information that can be used for semantic segmentation and instance segmentation. We also provide a comprehensive analysis of our annotation dataset from various aspects. We further measure the improvement of the state-of-the-art segmentation model after training it with our augmented dataset. The benefits of this augmented dataset are profound, with the potential to significantly enhance manga analysis algorithms and catalyze the novel development in digital art processing and cultural analytics. This annotation, named MangaSeg, is publicly available at https://huggingface.co/datasets/MS92/MangaSegmentation.
****
SeqMvRL: A Sequential Fusion Framework for Multi-view Representation Learning
Ren Wang, Haoliang Sun, Yuxiu Lin, Chuanhui Zuo, Yongshun Gong, Yilong Yin, Wenjia Meng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25822-25831
Multi-view representation learning integrates multiple observable views of an entity into a unified representation to facilitate downstream tasks. Current methods predominantly focus on distinguishing compatible components across views, followed by a single-step parallel fusion process. However, this parallel fusion is static in essence, overlooking potential conflicts among views and compromising representation ability. To address this issue, this paper proposes a novel Sequential fusion framework for Multi-view Representation Learning, termed SeqMvRL. Specifically, we model multi-view fusion as a sequential decision-making problem and construct a pairwise integrator (PI) and a next-view selector (NVS), which represent the environment and agent in reinforcement learning, respectively. PI merges the current fused feature with the selected view, while NVS is introduced to determine which view to fuse subsequently. By adaptively selecting the next optimal view for fusion based on the current fusion state, SeqMvRL thereby effectively reduces conflicts and enhances unified representation quality. Additionally, an elaborate novel reward function encourages the model to prioritize views that enhance the discriminability of the fused features. Experimental results demonstrate that SeqMvRL outperforms parallel fusion approaches in classification and clustering tasks.
****
Auto Cherry-Picker: Learning from High-quality Generative Data Driven by Language
Yicheng Chen, Xiangtai Li, Yining Li, Yanhong Zeng, Jianzong Wu, Xiangyu Zhao, Kai Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19952-19962
Diffusion models can generate realistic and diverse images, potentially facilitating data availability for data-intensive perception tasks. However, leveraging these models to boost performance on downstream tasks with synthetic data poses several challenges, including aligning with real data distribution, scaling synthetic sample volumes, and ensuring their quality. To bridge these gaps, we present Auto Cherry-Picker (ACP), a novel framework that generates high-quality cross-modality training samples at scale to augment perception and multi-modal traini

ng. ACP first uses LLMs to sample descriptions and layouts based on object combi nations from real data priors, eliminating the need for ground truth image capti ons or annotations. Next, we use an off-the-shelf controllable diffusion model t o generate multiple images. Then, the generated data are refined using a compreh ensively designed metric, Composite Layout and Image Score (CLIS), to ensure qua lity. Our customized synthetic high-quality samples boost performance in various scenarios, especially in addressing challenges associated with long-tailed dist ribution and imbalanced datasets. Experiment results on downstream tasks demonst rate that ACP can significantly improve the performance of existing models. In a ddition, we find a positive correlation between CLIS and performance gains in do wnstream tasks. This finding shows the potential for evaluation metrics as the r ole for various visual perception and MLLM tasks.
****

EnvGS: Modeling View-Dependent Appearance with Environment Gaussian
Tao Xie, Xi Chen, Zhen Xu, Yiman Xie, Yudong Jin, Yujun Shen, Sida Peng, Hujun B ao, Xiaowei Zhou; Proceedings of the Computer Vision and Pattern Recognition Con ference (CVPR), 2025, pp. 5742-5751
Reconstructing complex reflections in real-world scenes from 2D images is essent ial for achieving photorealistic novel view synthesis. Existing methods that uti lize environment maps to model reflections from distant lighting often struggle with high-frequency reflection details and fail to account for near-field reflec tions. In this work, we introduce EnvGS, a novel approach that employs a set of Gaussian primitives as an explicit 3D representation for capturing reflections o f environments. These environment Gaussian primitives are incorporated with base Gaussian primitives to model the appearance of the whole scene. To efficiently render these environment Gaussian primitives, we developed a ray-tracing-based r enderer that leverages the GPU's RT core for fast rendering. This allows us to j ointly optimize our model for high-quality reconstruction while maintaining real -time rendering speeds. Results from multiple real-world and synthetic datasets demonstrate that our method produces significantly more detailed reflections, ac hieving the best rendering quality in real-time novel view synthesis. The code i s available at https://zju3dv.github.io/envgs.
****

Provoking Multi-modal Few-Shot LVLM via Exploration-Exploitation In-Context Lear ning
Cheng Chen, Yunpeng Zhai, Yifan Zhao, Jinyang Gao, Bolin Ding, Jia Li; Proceedin gs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3 826-3835
In-context learning (ICL), a predominant trend in instruction learning, aims at enhancing the performance of large language models by providing clear task guida nce and examples, improving their capability in task understanding and execution . This paper investigates ICL on Large Vision-Language Models (LVLMs) and explor es the policies of multi-modal demonstration selection. Existing research effort s in ICL face significant challenges: First, they rely on pre-defined demonstrat ions or heuristic selecting strategies based on human intuition, which are usual ly inadequate for covering diverse task requirements, leading to sub-optimal sol utions; Second, individually selecting each demonstration fails in modeling the interactions between them, resulting in information redundancy. Unlike these pre vailing efforts, we propose a new exploration-exploitation reinforcement learnin g framework, which explores policies to fuse multi-modal information and adaptiv ely select adequate demonstrations as an integrated whole. The framework allows LVLMs to optimize themselves by continually refining their demonstrations throug h self-exploration, enabling the ability to autonomously identify and generate t he most effective selection policies for in-context learning. Experimental resul ts verify that our approach achieves significant performance improvements on fou r Visual Question-Answering (VQA) datasets, demonstrating its effectiveness in e nhancing the generalization capability of few-shot LVLMs.
****

BadToken: Token-level Backdoor Attacks to Multi-modal Large Language Models
Zenghui Yuan, Jiawen Shi, Pan Zhou, Neil Zhenqiang Gong, Lichao Sun; Proceedings

of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29927-29936
Multi-modal large language models (MLLMs) extend large language models (LLMs) to process multi-modal information, enabling them to generate responses to image-text inputs. MLLMs have been incorporated into diverse multi-modal applications, such as autonomous driving and medical diagnosis, via plug-and-play without fine-tuning. This deployment paradigm increases the vulnerability of MLLMs to backdoor attacks. However, existing backdoor attacks against MLLMs achieve limited effectiveness and stealthiness. In this work, we propose BadToken, the first token-level backdoor attack to MLLMs. BadToken introduces two novel backdoor behaviors: Token-substitution and Token-addition, which enable flexible and stealthy attacks by making token-level modifications to the original output for backdoored inputs. We formulate a general optimization problem that considers the two backdoor behaviors to maximize the attack effectiveness. We evaluate BadToken on two open-source MLLMs and various tasks. Our results show that our attack maintains the model's utility while achieving high attack success rates and stealthiness. We also show the real-world threats of BadToken in two scenarios, i.e., autonomous driving and medical diagnosis. Furthermore, we consider defenses including fine-tuning and input purification. Our results highlight the threat of our attack.
****
ReRAW: RGB-to-RAW Image Reconstruction via Stratified Sampling for Efficient Object Detection on the Edge
Radu Berdan, Beril Besbinar, Christoph Reinders, Junji Otsuka, Daisuke Iso; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11833-11843
Edge-based computer vision models running on compact, resource-limited devices benefit greatly from using unprocessed, detail-rich RAW sensor data instead of processed RGB images. Training these models, however, necessitates large labeled RAW datasets, which are costly and often impractical to obtain. Thus, converting existing labeled RGB datasets into sensor-specific RAW images becomes crucial for effective model training. In this paper, we introduce ReRAW, an RGB-to-RAW conversion model that achieves state-of-the-art reconstruction performance across five diverse RAW datasets. This is accomplished through ReRAW's novel multi-head architecture predicting RAW image candidates in gamma space. The performance is further boosted by a stratified sampling-based training data selection heuristic, which helps the model better reconstruct brighter RAW pixels. We finally demonstrate that pretraining compact models on a combination of high-quality synthetic RAW datasets (such as generated by ReRAW) and ground-truth RAW images for downstream tasks like object detection, outperforms both standard RGB pipelines, and RAW fine-tuning of RGB-pretrained models for the same task.
****
VLMs-Guided Representation Distillation for Efficient Vision-Based Reinforcement Learning
Haoran Xu, Peixi Peng, Guang Tan, Yiqian Chang, Luntong Li, Yonghong Tian; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29534-29544
Vision-based Reinforcement Learning (VRL) attempts to establish associations between visual inputs and optimal actions through interactions with the environment. Given the high-dimensional and complex nature of visual data, it becomes essential to learn policy upon high-quality state representation. To this end, existing VRL methods primarily rely on interaction-collected data, combined with self-supervised auxiliary tasks. However, two key challenges remain: limited data samples and a lack of task-relevant semantic constraints. To tackle this, we propose DGC, a method that distills guidance from Visual Language Models (VLMs) alongside self-supervised learning into a compact VRL agent. Notably, we leverage the state representation capabilities of VLMs, rather than their decision-making abilities. Within DGC, a novel prompting-reasoning pipeline is designed to convert historical observations and actions into usable supervision signals, enabling semantic understanding within the compact visual encoder. By leveraging these distilled semantic representations, the VRL agent achieves significant improvements

in the sample efficiency. Extensive experiments on the Carla benchmark demonstrate our state-of-the-art performance.
****

MonoDGP: Monocular 3D Object Detection with Decoupled-Query and Geometry-Error Priors

Fanqi Pu, Yifan Wang, Jiru Deng, Wenming Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6520-6530

Perspective projection has been extensively utilized in monocular 3D object detection methods. It introduces geometric priors from 2D bounding boxes and 3D object dimensions to reduce the uncertainty of depth estimation. However, due to errors originating from the object's visual surface, the bounding box height often fails to represent the actual central height, which undermines the effectiveness of geometric depth. Direct prediction for the projected height unavoidably results in a loss of 2D priors, while multi-depth prediction with complex branches does not fully leverage geometric depth. This paper presents a Transformer-based monocular 3D object detection method called MonoDGP, which adopts perspective-invariant geometry errors to modify the projection formula. We also try to systematically discuss and explain the mechanisms and efficacy behind geometry errors, which serve as a simple but effective alternative to multi-depth prediction. Additionally, MonoDGP decouples the depth-guided decoder and constructs a 2D decoder only dependent on visual features, providing 2D priors and initializing object queries without the disturbance of 3D detection. To further optimize and fine-tune input tokens of the transformer decoder, we also introduce a Region Segmentation Head (RSH) that generates enhanced features and segment embeddings. Our monocular method demonstrates state-of-the-art performance on the KITTI benchmark without extra data. Code is available at https://github.com/PuFanqi23/MonoDGP.
****

NeISF++: Neural Incident Stokes Field for Polarized Inverse Rendering of Conductors and Dielectrics

Chenhao Li, Taishi Ono, Takeshi Uemori, Sho Nitta, Hajime Mihara, Alexander Gatto, Hajime Nagahara, Yusuke Moriuchi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26493-26503

Recent inverse rendering methods have improved shape, material, and illumination reconstruction using polarization cues. However, they only support dielectrics, ignoring conductors, which are common in everyday life. Since conductors and dielectrics have different reflection properties, using previous dielectrics-based methods will lead to obvious errors. In addition, conductors are glossy, which may cause strong specular reflection and is hard to reconstruct. To solve the above issues, we propose NeISF++, an inverse rendering pipeline that supports conductors and dielectrics. The key ingredient for our proposal is a general pBRDF that describes both conductors and dielectrics. As for the strong specular reflection problem, we propose a novel geometry initialization method using DoLP images. This physical cue is invariant to intensities and thus robust to strong specular reflections. Experimental results on our synthetic and real datasets show that our method surpasses the existing polarized inverse rendering methods for geometry and material decomposition as well as downstream tasks like relighting.
****

HunyuanPortrait: Implicit Condition Control for Enhanced Portrait Animation

Zunnan Xu, Zhentao Yu, Zixiang Zhou, Jun Zhou, Xiaoyu Jin, Fa-ting Hong, Xiaozhong Ji, Junwei Zhu, Chengfei Cai, Shiyu Tang, Qin Lin, Xiu Li, Qinglin Lu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15909-15919

We introduce HunyuanPortrait, a diffusion-based condition control method that employs implicit representations for highly controllable and lifelike portrait animation. Given a single portrait image as an appearance reference and video clips as driving templates, HunyuanPortrait can animate the character in the reference image by the facial expression and head pose of the driving videos. In our framework, we utilize pre-trained encoders to achieve the decoupling of portrait motion information and identity in videos. To do so, implicit representation is adopted to encode motion information and is employed as control signals in the ani

mation phase. By leveraging the power of stable video diffusion as the main building block, we carefully design adapter layers to inject control signals into the denoising unet through attention mechanisms. These bring spatial richness of details and temporal consistency. HunyuanPortrait also exhibits strong generalization performance, which can effectively disentangle appearance and motion under different image styles. Our framework outperforms existing methods, demonstrating superior temporal consistency and controllability. Our project is available at https://kkakkkka.github.io/HunyuanPortrait.
****

Flexible Group Count Enables Hassle-Free Structured Pruning
Jiamu Zhang, Shaochen Zhong, Andrew Ye, Zirui Liu, Sebastian Zhao, Kaixiong Zhou, Li Li, Soo-Hyun Choi, Rui Chen, Xia Hu, Shuai Xu, Vipin Chaudhary; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4807-4818
Densely structured pruning methods -- which generate pruned models in a fully dense format, allowing immediate compression benefits without additional demands -- are evolving owing to their practical significance. Traditional techniques in this domain mainly revolve around coarser granularities, such as filter pruning, thereby limiting their performance due to restricted pruning freedom.Recent advancements in Grouped Kernel Pruning (GKP) have enabled the utilization of finer granularity while maintaining the densely structured format. We observed that existing GKP methods often introduce dynamic operations to different aspects of their procedures, where many were done so at the cost of adding complications and/or imposing limitations -- e.g., requiring an expensive mixture of clustering schemes; or having dynamic pruning rates and sizes among groups, which lead to reliance on custom architecture support for its pruned models.In this work, we argue the best practice to introduce such dynamic operation to GKP is to make `Conv2d(groups)` (a.k.a. group count) flexible under an integral optimization, leveraging its ideal alignment with the infrastructure support of Grouped Convolution. Pursuing such direction, we present a one-shot, post-train, data-agnostic GKP method that is more performant, adaptive, and efficient than its predecessors; while simultaneously being a lot more user-friendly with little-to-no hyper-parameter tuning or handcrafted criteria required.
****

EasyCraft: A Robust and Efficient Framework for Automatic Avatar Crafting
Suzhen Wang, Weijie Chen, Wei Zhang, Minda Zhao, Lincheng Li, Rongsheng Zhang, Zhipeng Hu, Xin Yu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5581-5591
Character customization, or 'face crafting,' is a vital feature in role-playing games (RPGs), enhancing player engagement by enabling the creation of personalized avatars. Existing automated methods often struggle with generalizability across diverse game engines due to their reliance on the intermediate constraints of specific image domain and typically support only one type of input, either text or image. To overcome these challenges, we introduce EasyCraft, an innovative end-to-end feedforward framework that automates character crafting by uniquely supporting both text and image inputs. Our approach employs a translator capable of converting facial images of any style into crafting parameters. We first establish a unified feature distribution in the translator's image encoder through self-supervised learning on a large-scale dataset, enabling photos of any style to be embedded into a unified feature representation. Subsequently, we map this unified feature distribution to crafting parameters specific to a game engine, a process that can be easily adapted to most game engines and thus enhances EasyCraft's generalizability. By integrating text-to-image techniques with our translator, EasyCraft also facilitates precise, text-based character crafting. EasyCraft's ability to integrate diverse inputs significantly enhances the versatility and accuracy of avatar creation. Extensive experiments on two RPG games demonstrate the effectiveness of our method, achieving state-of-the-art results and facilitating adaptability across various avatar engines.
****

MeshArt: Generating Articulated Meshes with Structure-Guided Transformers

Daoyi Gao, Yawar Siddiqui, Lei Li, Angela Dai; Proceedings of the Computer Visio n and Pattern Recognition Conference (CVPR), 2025, pp. 618-627
Articulated 3D object generation is fundamental for creating realistic, function al, and interactable virtual assets which are not simply static. We introduce Me shArt, a hierarchical transformer-based approach to generate articulated 3D mesh es with clean, compact geometry, reminiscent of human-crafted 3D models. We appr oach articulated mesh generation in a part-by-part fashion across two stages.Fir st, we generate a high-level articulation-aware object structure; then, based on this structural information, we synthesize each part's mesh faces. Key to our a pproach is modeling both articulation structures and part meshes as sequences of quantized triangle embeddings, leading to a unified hierarchical framework with transformers for autoregressive generation.Object part structures are first gen erated as their bounding primitives and articulation modes; a second transformer , guided by these articulation structures, then generates each part's mesh trian gles.To ensure coherency among generated parts, we introduce structure-guided co nditioning that also incorporates local part mesh connectivity.MeshArt shows si gnificant improvements over state of the art, with 57.1% improvement in structur e coverage and a 209-point improvement in mesh generation FID.
****

Non-Natural Image Understanding with Advancing Frequency-based Vision Encoders
Wang Lin, QingSong Wang, Yueying Feng, Shulei Wang, Tao Jin, Zhou Zhao, Fei Wu, Chang Yao, Jingyuan Chen; Proceedings of the Computer Vision and Pattern Recogni tion Conference (CVPR), 2025, pp. 29756-29766
Large language models (LLMs) have significantly enhanced cross-modal understandi ng capabilities by integrating visual encoders with textual embeddings, giving r ise to multimodal large language models (MLLMs). However, these models struggle with non-natural images such as geometric and charts, particularly in fields lik e education and finance. Despite efforts to collect datasets and fine-tune the M LLMs, the gap with natural image understanding is still evident, and the cost of collecting large and diverse non-natural image datasets is high. To address thi s, we analyzed the limitations of transformer-based vision encoders(ViT) within existing MLLMs from a frequency perspective. Studies have shown that ViT models are less effective at capturing high-frequency information, impairing their abil ity to capture elements like points, lines, and angles in non-natural images. In response, we introduced FM-ViT, a frequency-modulated vision encoder that utili zes Fourier decomposition to extract high and low frequency components from self -attention features and re-weight them during tuning to non-natural images. In a ddition, we combine the features of CNN models with FM-ViT and propose EDGE, an MLLM with enhanced graphical encoders tailored for understanding non-natural ima ges. Extensive experiments have confirmed the effectiveness of our FM-ViT and ED GE in 4 types.
****

Generative Multimodal Pretraining with Discrete Diffusion Timestep Tokens
Kaihang Pan, Wang Lin, Zhongqi Yue, Tenglong Ao, Liyu Jia, Wei Zhao, Juncheng Li , Siliang Tang, Hanwang Zhang; Proceedings of the Computer Vision and Pattern Re cognition Conference (CVPR), 2025, pp. 26136-26146
Recent endeavors in Multimodal Large Language Models (MLLMs) aim to unify visual comprehension and generation by combining LLM and diffusion models, the state-o f-the-art in each task, respectively. Existing approaches rely on spatial visual tokens, where image patches are encoded and arranged according to a spatial ord er (e.g., raster scan). However, we show that spatial tokens lack the recursive structure inherent to languages, hence form an impossible language for LLM to ma ster. In this paper, we build a proper visual language by leveraging diffusion t imesteps to learn discrete, recursive visual tokens. Our proposed tokens recursi vely compensate for the progressive attribute loss in noisy images as timesteps increase, enabling the diffusion model to reconstruct the original image at any timestep. This approach allows us to effectively integrate the strengths of LLMs in autoregressive reasoning and diffusion models in precise image generation, a chieving seamless multimodal comprehension and generation within a unified frame work. Extensive experiments show that we achieve a new SOTA for multimodal comp

rehension and generation simultaneously compared with other MLLMs. Project Page: https://DDT-LLaMA.github.io/.
****

SplatFlow: Self-Supervised Dynamic Gaussian Splatting in Neural Motion Flow Field for Autonomous Driving
Su Sun, Cheng Zhao, Zhuoyang Sun, Yingjie Victor Chen, Mei Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27487-27496
Most existing Dynamic Gaussian Splatting methods for complex dynamic urban scenarios rely on accurate object-level supervision from expensive manual labeling, limiting their scalability in real-world applications. In this paper, we introduce SplatFlow, a Self-Supervised Dynamic Gaussian Splatting within Neural Motion Flow Fields (NMFF) to learn 4D space-time representations without requiring tracked 3D bounding boxes, enabling accurate dynamic scene reconstruction and novel view RGB/depth/flow synthesis. SplatFlow designs a unified framework to seamlessly integrate time-dependent 4D Gaussian representation within NMFF, where NMFF is a set of implicit functions to model temporal motions of both LiDAR points and Gaussians as continuous motion flow fields. Leveraging NMFF, SplatFlow effectively decomposes static background and dynamic objects, representing them with 3D and 4D Gaussian primitives, respectively. NMFF also models the correspondences of each 4D Gaussian across time, which aggregates temporal features to enhance cross-view consistency of dynamic components. SplatFlow further improves dynamic object identification by distilling features from 2D foundation models into 4D space-time representation. Comprehensive evaluations conducted on the Waymo and KITTI Datasets validate SplatFlow's state-of-the-art (SOTA) performance for both image reconstruction and novel view synthesis in dynamic urban scenarios.
****

Zero-shot 3D Question Answering via Voxel-based Dynamic Token Compression
Hsiang-Wei Huang, Fu-Chen Chen, Wenhao Chai, Che-Chun Su, Lu Xia, Sanghun Jung, Cheng-Yen Yang, Jenq-Neng Hwang, Min Sun, Cheng-Hao Kuo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19424-19434
Recent advancements in 3D Large Multi-modal Models (3D-LMMs) have driven significant progress in 3D question answering. However, recent multi-frame Vision-Language Models (VLMs) demonstrate superior performance compared to 3D-LMMs on 3D question answering tasks, largely due to the greater scale and diversity of available 2D image data in contrast to the more limited 3D data. Multi-frame VLMs, although achieving superior performance, suffer from the difficulty of retaining all the detailed visual information in the 3D scene while limiting the number of visual tokens. Common methods such as token pooling, reduce visual token usage but often lead to information loss, impairing the model's ability to preserve visual details essential for 3D question answering tasks. To address this, we propose voxel-based Dynamic Token Compression (DTC), which combines 3D spatial priors and visual semantics to achieve over 90% reduction in visual tokens usage for current multi-frame VLMs. Our method maintains performance comparable to state-of-the-art models on 3D question answering benchmarks including OpenEQA and ScanQA, demonstrating its effectiveness.
****

AesthetiQ: Enhancing Graphic Layout Design via Aesthetic-Aware Preference Alignment of Multi-modal Large Language Models
Sohan Patnaik, Rishabh Jain, Balaji Krishnamurthy, Mausoom Sarkar; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23701-23711
Visual layouts are essential in graphic design fields such as advertising, posters, and web interfaces. The application of generative models for content-aware layout generation has recently gained traction. However, these models fail to understand the contextual aesthetic requirements of layout design and do not align with human-like preferences, primarily treating it as a prediction task without considering the final rendered output. To overcome these problems, we offer Aesthetic-Aware Preference Alignment (AAPA), a novel technique to train a Multi-modal Large Language Model (MLLM) for layout prediction that uses MLLM's aesthetic p

references for Direct Preference Optimization over graphic layouts. We propose a data filtering protocol utilizing our layout-quality heuristics for AAPA to ensure training happens on high-quality layouts. Additionally, we introduce a novel evaluation metric that uses another MLLM to compute the win rate of the generated layout against the ground-truth layout based on aesthetics criteria. We also demonstrate the applicability of AAPA for MLLMs of varying scales (1B to 8B parameters) and LLM families (Qwen, Phi, InternLM). By conducting thorough qualitative and quantitative analyses, we verify the efficacy of our approach on two challenging benchmarks - Crello and Webui, showcasing 17%, and 16% improvement over current State-of-The-Art methods, thereby highlighting the potential of MLLMs in aesthetic-aware layout generation.

****

Enhanced then Progressive Fusion with View Graph for Multi-View Clustering

Zhibin Dong, Meng Liu, Siwei Wang, Ke Liang, Yi Zhang, Suyuan Liu, Jiaqi Jin, Xinwang Liu, En Zhu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15518-15527

Multi-view clustering aims to improve clustering accuracy by effectively integrating complementary information from multiple perspectives. However, existing methods often encounter challenges such as feature conflicts between views and insufficient enhancement of individual view features, which hinder clustering performance. To address these challenges, we propose a novel framework, EPFMVC, which integrates feature enhancement with progressive fusion to more effectively align multi-view data. Specifically, we introduce two key innovations: (1) a Feature Channel Attention Encoder (FCAencoder), which adaptively enhances the most discriminative features in each view, and (2) a View Graph-based Progressive Fusion Mechanism, which constructs a view graph using optimal transport (OT) distance to progressively fuse similar views while minimizing inter-view conflicts. By leveraging multi-head attention, the fusion process gradually integrates complementary information, ensuring more consistent and robust shared representations. These innovations enable superior representation learning and effective fusion across views. Extensive experiments demonstrate that our method significantly outperforms state-of-the-art techniques, achieving notable improvements in multi-view clustering tasks across various datasets and evaluation metrics.

****

FINECAPTION: Compositional Image Captioning Focusing on Wherever You Want at Any Granularity

Hang Hua, Qing Liu, Lingzhi Zhang, Jing Shi, Soo Ye Kim, Zhifei Zhang, Yilin Wang, Jianming Zhang, Zhe Lin, Jiebo Luo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24763-24773

The advent of large Vision-Language Models (VLMs) has significantly advanced multimodal tasks, enabling more sophisticated and accurate integration of visual and textual information across various applications, including image and video captioning, visual question answering, and cross-modal retrieval.Despite their superior capabilities, VLMs still struggle with fine-grained compositional image region descriptions. Specifically, they have difficulty recognizing arbitrary segmentation masks as referential inputs, interpreting compositional aspect instructions for referencing, and precisely describing the compositional aspects of a region. However, compositionality--the ability to understand and generate novel combinations of known visual and textual components--is critical for facilitating coherent reasoning and understanding across modalities in VLMs. To address this issue, we propose OpenCompositionCap, a new dataset for multi-grained region compositional image captioning that distinguishes itself from prior works by introducing the new task of compositional aspect-aware regional image captioning. To support this endeavor, we also introduce a new VLM model, FineCaption. The empirical results illustrate the effectiveness of our proposed model compared with other strong VLMs. In addition, we analyze the capabilities of current VLMs in recognizing various visual prompts for compositional region image captioning, highlighting areas for improvement in VLM design and training.

****

Adaptive Non-Uniform Timestep Sampling for Accelerating Diffusion Model Training

Myunsoo Kim, Donghyeon Ki, Seong-Woong Shim, Byung-Jun Lee; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2513-2522

As a highly expressive generative model, diffusion models have demonstrated exceptional success across various domains, including image generation, natural language processing, and combinatorial optimization. However, as data distributions grow more complex, training these models to convergence becomes increasingly computationally intensive. While diffusion models are typically trained using uniform timestep sampling, our research shows that the variance in stochastic gradients varies significantly across timesteps, with high-variance timesteps becoming bottlenecks that hinder faster convergence. To address this issue, we introduce a non-uniform timestep sampling method that prioritizes these more critical timesteps. Our method tracks the impact of gradient updates on the objective for each timestep, adaptively selecting those most likely to minimize the objective effectively. Experimental results demonstrate that this approach not only accelerates the training process, but also leads to improved performance at convergence. Furthermore, our method shows robust performance across various datasets, scheduling strategies, and diffusion architectures, outperforming previously proposed timestep sampling and weighting heuristics that lack this degree of robustness.

****

Chebyshev Attention Depth Permutation Texture Network with Latent Texture Attribute Loss

Ravishankar Evani, Deepu Rajan, Shangbo Mao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23423-23432

Despite recent advances in deep texture recognition, existing methods still lack representational diversity and struggle to capture and preserve discriminative cues across stages of representation hierarchies. Moreover, many rely on loss formulations that prioritize recognition accuracy while overlooking spatial coherence and statistical consistency in the feature space. To address these issues, we propose three key innovations: Stochastic Local Texture Masking (SLTM), a regularization strategy that randomly occludes small texture patches to promote the learning of broader spatial and contextual dependencies; the Chebyshev Attention Depth Permutation Texture Network (CAPTN), a novel architecture that learns expressive and persistent Latent Texture Attribute (LTA) representations. CAPTN integrates a Texture Frequency Attention (TFA) module that generates LTAs and enables frequency-aware interpretability, a Dual Depth Permutation (D2P) module to expose complementary channel adjacency patterns, and Learnable Chebyshev Polynomials (LCPs) to model high-order orderless LTA transformations via recursive Chebyshev basis expansion; and a Latent Texture Attribute Loss that jointly optimizes classification accuracy, statistical alignment, and spatial fidelity. CAPTN supports end-to-end training without relying on fine-tuned CNN backbones and achieves state-of-the-art performance on several texture and material recognition benchmarks. (Code: https://github.com/RavishankarEvani/CAPTN)

****

Explainable Saliency: Articulating Reasoning with Contextual Prioritization

Nuo Chen, Ming Jiang, Qi Zhao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9601-9610

Deep saliency models, which predict what parts of an image capture our attention, are often like black boxes. This limits their use, especially in areas where understanding why a model makes a decision is crucial. Our research tackles this challenge by developing an explainable saliency (XSal) model that not only identifies what is important in an image, but also explains its choices in a way that makes sense to humans. We achieve this by using vision-language models to reason about images and by focusing the model's attention on the most crucial information using a contextual prioritization mechanism. Unlike prior approaches that rely on fixation descriptions or soft-attention based semantic aggregation, our method directly models the reasoning steps involved in saliency prediction, generating selectively prioritized explanations clarify why specific regions are prioritized. Comprehensive evaluations demonstrate the effectiveness of our model in generating high-quality saliency maps and coherent, contextually relevant explanations. This research is a step towards more transparent and trustworthy AI sys

tems that can help us understand and navigate the world around us.
****

Decentralized Diffusion Models
David McAllister, Matthew Tancik, Jiaming Song, Angjoo Kanazawa; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23323-23333
Large-scale AI model training divides work across thousands of GPUs then synchronizes gradients across them at each step. This incurs a significant network burden that only centralized, monolithic clusters can support, driving up infrastructure costs and straining power systems. We propose Decentralized Diffusion Models, a scalable framework to distribute diffusion model training across independent clusters or datacenters by eliminating the dependence on a centralized, high-bandwidth networking fabric. Our method trains a set of expert diffusion models over partitions of the dataset, each in full isolation from one another. At inference time, they ensemble through a lightweight router. We show that this ensemble collectively optimizes the same objective as a single model trained over the whole dataset. This means we can divide the training burden among a number of "compute islands," lowering infrastructure costs and improving resilience to localized GPU failures. Decentralized diffusion models empower researchers to take advantage of smaller, more cost-effective and more readily available compute like on-demand GPU nodes rather than central integrated systems. We conduct extensive experiments on ImageNet and LAION Aesthetics, showing that decentralized diffusion models FLOP-for-FLOP outperform standard diffusion models. We finally scale our approach to 24 billion parameters, demonstrating that high-quality diffusion models can now be trained with just eight individual GPU nodes in less than a week.
****

AnyEdit: Mastering Unified High-Quality Image Editing for Any Idea
Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, Yueting Zhuang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26125-26135
Instruction-based image editing aims to modify specific image elements with natural language instructions. However, current models in this domain often struggle to execute complex user instructions accurately, as they are trained on low-quality data with limited editing types. We present AnyEdit, a comprehensive multi-modal instruction editing dataset, comprising 2.5 million high-quality editing pairs spanning over 20 editing types and five domains. We ensure the diversity and quality of the AnyEdit collection through three aspects: initial data diversity, adaptive editing process, and automated selection of editing results. Using the dataset, we further train a novel AnyEdit Stable Diffusion with task-aware routing and learnable task embedding for unified image editing. Comprehensive experiments on three benchmark datasets show that AnyEdit consistently boosts the performance of diffusion-based editing models. This presents prospects for developing instruction-driven image editing models that support human creativity.
****

Compass Control: Multi Object Orientation Control for Text-to-Image Generation
Rishubh Parihar, Vaibhav Agrawal, Sachidanand VS, Venkatesh Babu Radhakrishnan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2791-2801
Existing approaches for controlling text-to-image diffusion models, while powerful, do not allow for explicit 3D object-centric control, such as precise control of object orientation. In this work, we address the problem of multi-object orientation control in text-to-image diffusion models. This enables the generation of diverse multi-object scenes with precise orientation control for each object. The key idea is to condition the diffusion model with a set of orientation-aware compass tokens, one for each object, along with text tokens. A light-weight encoder network predicts these compass tokens taking object orientation as the input. The model is trained on a synthetic dataset of procedurally generated scenes, each containing one or two 3D-assets on a plain background. However, direct training this framework results in poor orientation control as well as leads to en

tanglement among objects. To mitigate this, we intervene in the generation proce
ss and constrain the cross-attention maps of each compass token to its correspon
ding object regions. The trained model is able to achieve precise orientation co
ntrol for a) complex objects not seen during training, and b) multi-object scene
s with more than two objects, indicating strong generalization capabilities. Fur
ther, when combined with personalization methods, our method precisely controls
the orientation of the new object in diverse contexts. Our method achieves state
-of-the-art orientation control and text alignment quantified with extensive eva
luations and a user study.
****

Correcting Deviations from Normality: A Reformulated Diffusion Model for Multi-C
lass Unsupervised Anomaly Detection
Farzad Beizaee, Gregory A. Lodygensky, Christian Desrosiers, Jose Dolz; Proceedi
ngs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp.
19088-19097
Recent advances in diffusion models have spurred research into their application
 for Reconstruction-based unsupervised anomaly detection. However, these methods
 may struggle with maintaining structural integrity and recovering the anomaly-f
ree content of abnormal regions, especially in multi-class scenarios. Furthermor
e, diffusion models are inherently designed to generate images from pure noise a
nd struggle to selectively alter anomalous regions of an image while preserving
normal ones. This leads to potential degradation of normal regions during recons
truction, hampering the effectiveness of anomaly detection. This paper introduc
es a reformulation of the standard diffusion model geared toward selective regio
n alteration, allowing the accurate identification of anomalies. By modeling ano
malies as noise in the latent space, our proposed Deviation correction diffusion
 (DeCo-Diff) model preserves the normal regions and encourages transformations e
xclusively on anomalous areas. This selective approach enhances the reconstructi
on quality, facilitating effective unsupervised detection and localization of an
omaly regions. Comprehensive evaluations demonstrate the superiority of our meth
od in accurately identifying and localizing anomalies in complex images, with pi
xel-level AUPRC improvements of 11-14% over state-of-the-art models on well-know
n anomaly detection datasets. The code is available at https://github.com/farzad
-bz/DeCo-Diff
****

Continuous 3D Perception Model with Persistent State
Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, Angjoo Kanazaw
a; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR),
 2025, pp. 10510-10522
We present a unified framework capable of solving a broad range of 3D tasks. Our
 approach features a stateful recurrent model that continuously updates its stat
e representation with each new observation. Given a stream of images, this evolv
ing state can be used to generate metric-scale pointmaps (per-pixel 3D points) f
or each new input in an online fashion. These pointmaps reside within a common c
oordinate system, and can be accumulated into a coherent, dense scene reconstruc
tion that updates as new images arrive. Our model, called CUT3R (Continuous Upda
ting Transformer for 3D Reconstruction), captures rich priors of real-world scen
es: not only can it predict accurate pointmaps from image observations, but it c
an also infer unseen regions of the scene by probing at virtual, unobserved view
s. Our method is simple yet highly flexible, naturally accepting varying length
of images that may be either video streams or unordered photo collections, conta
ining both static and dynamic content. We evaluate our method on various 3D/4D t
asks and demonstrate competitive or state-of-the-art performance in each. Projec
t page: https://cut3r.github.io/.
****

LP-Diff: Towards Improved Restoration of Real-World Degraded License Plate
Haoyan Gong, Zhenrong Zhang, Yuzheng Feng, Anh Nguyen, Hongbin Liu; Proceedings
of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1783
1-17840
License plate (LP) recognition is crucial in intelligent traffic management syst

ems. However, factors such as long distances and poor camera quality often lead to severe degradation of captured LP images, posing challenges to accurate recognition. The design of License Plate Image Restoration (LPIR) methods frequently relies on synthetic degraded data, which limits their effectiveness on real-world severely degraded LP images. To address this issue, we introduce the first paired LPIR dataset collected in real-world scenarios, named MDLP, including 10,245 pairs of multi-frame severely degraded LP images and their corresponding clear images. To better restore severely degraded LP, we propose a novel Diffusion-based network, called LP-Diff, to tackle real-world LPIR tasks. Our approach incorporates (1) an Inter-frame Cross Attention Module to fuse temporal information across multiple frames, (2) a Texture Enhancement Module to restore texture information in degraded images, and (3) a Dual-Pathway Fusion Module to select effective features from both channel and spatial dimensions. Extensive experiments demonstrate the reliability of our dataset for model training and evaluation. Our proposed LP-Diff consistently outperforms other state-of-the-art image restoration methods on real-world LPIR tasks. The dataset and code are available at https://github.com/haoyGONG/LP-Diff.

****

Unleashing the Potential of Consistency Learning for Detecting and Grounding Multi-Modal Media Manipulation

Yiheng Li, Yang Yang, Zichang Tan, Huan Liu, Weihua Chen, Xu Zhou, Zhen Lei; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9242-9252

To tackle the threat of fake news, the task of detecting and grounding multi-modal media manipulation (DGM4) has received increasing attention. However, most state-of-the-art methods fail to explore the fine-grained consistency within local content, usually resulting in an inadequate perception of detailed forgery and unreliable results. In this paper, we propose a novel approach named Contextual-Semantic Consistency Learning (CSCL) to enhance the fine-grained perception ability of forgery for DGM^4. Two branches for image and text modalities are established, each of which contains two cascaded decoders, i.e., Contextual Consistency Decoder (CCD) and Semantic Consistency Decoder (SCD), to capture within-modality contextual consistency and across-modality semantic consistency, respectively. Both CCD and SCD adhere to the same criteria for capturing fine-grained forgery details. To be specific, each module first constructs consistency features by leveraging additional supervision from the heterogeneous information of each token pair. Then, the forgery-aware reasoning or aggregating is adopted to deeply seek forgery cues based on the consistency features. Extensive experiments on DGM4 datasets prove that CSCL achieves new state-of-the-art performance, especially for the results of grounding manipulated content. Codes and weights are avaliable at https://github.com/liyih/CSCL.

****

DNF: Unconditional 4D Generation with Dictionary-based Neural Fields

Xinyi Zhang, Naiqi Li, Angela Dai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26047-26056

While remarkable success has been achived through diffusion-based 3D generative models for shapes, 4D generative modeling remains challenging due to the complexity of object deformations over time. We propose DNF, a new 4D representation for unconditional generative modeling that efficiently models deformable shapes with disentangled shape and motion while capturing high-fidelity details in the deforming objects. To achieve this, we propose a dictionary learning approach to disentangle 4D motion from shape as neural fields.Both shape and motion are represented as learned latent spaces, where each deformable shape is represented by its shape and motion global latent codes, shape-specific coefficient vectors, and shared dictionary information. This captures both shape-specific detail and global shared information in the learned dictionary. Our dictionary-based representation well balances fidelity, contiguity and compression -- combined with a transformer-based diffusion model, our method is able to generate effective, high-fidelity 4D animations.

****

ARM: Appearance Reconstruction Model for Relightable 3D Generation
Xiang Feng, Chang Yu, Zoubin Bi, Yintong Shang, Feng Gao, Hongzhi Wu, Kun Zhou, Chenfanfu Jiang, Yin Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21425-21437
Recent image-to-3D reconstruction models have greatly advanced geometry generation, but they still struggle to faithfully generate realistic appearance. To address this, we introduce ARM, a novel method that reconstructs high-quality 3D meshes and realistic appearance from sparse-view images. The core of ARM lies in decoupling geometry from appearance, processing appearance within the UV texture space. Unlike previous methods, ARM improves texture quality by explicitly back-projecting measurements onto the texture map and processing them in a UV space module with a global receptive field. To resolve ambiguities between material and illumination in input images, ARM introduces a material prior that encodes semantic appearance information, enhancing the robustness of appearance decomposition. Trained on just 8 H100 GPUs, ARM outperforms existing methods both quantitatively and qualitatively. Our project page is available at https://arm-aigc.github.io/.
****

VideoGEM: Training-free Action Grounding in Videos
Felix Vogel, Walid Bousselham, Anna Kukleva, Nina Shvetsova, Hilde Kuehne; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3374-3383
Vision-language foundation models have shown impressive capabilities across various zero-shot tasks, including training-free localization and grounding, primarily focusing on localizing objects in images. However, leveraging those capabilities to localize actions and events in videos is challenging, as actions have less physical outline and are usually described by higher-level concepts.In this work, we propose VideoGEM, the first training-free spatial action grounding method based on pretrained image- and video-language backbones. Namely, we adapt the self-self attention formulation of GEM to spatial activity grounding. We observe that high-level semantic concepts, such as actions, usually emerge in the higher layers of the image- and video-language models. We, therefore, propose a layer weighting in the self-attention path to prioritize higher layers. Additionally, we introduce a dynamic weighting method to automatically tune layer weights to capture each layer's relevance to a specific prompt. Finally, we introduce a prompt decomposition, processing action, verb, and object prompts separately, resulting in a better spatial localization of actions. We evaluate the proposed approach on three image- and video-language backbones, CLIP, OpenCLIP, and ViCLIP, and on four video grounding datasets, V-HICO, DALY, YouCook-Interactions, and GroundingYouTube, showing that the proposed training-free approach is able to outperform current trained state-of-the-art approaches for spatial video grounding.
****

FilmComposer: LLM-Driven Music Production for Silent Film Clips
Zhifeng Xie, Qile He, Youjia Zhu, Qiwei He, Mengtian Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13519-13528
In this work, we implement music production for silent film clips using LLM-driven method. Given the strong professional demands of film music production, we propose the FilmComposer, simulating the actual workflows of professional musicians. FilmComposer is the first to combine large generative models with a multi-agent approach, leveraging the advantages of both waveform music and symbolic music generation. Additionally, FilmComposer is the first to focus on the three core elements of music production for film--audio quality, musicality, and musical development--and introduces various controls, such as rhythm, semantics, and visuals, to enhance these key aspects. Specifically, FilmComposer consists of the visual processing module, rhythm-controllable MusicGen, and multi-agent assessment, arrangement and mix. In addition, our framework can seamlessly integrate into the actual music production pipeline and allows user intervention in every step, providing strong interactivity and a high degree of creative freedom. Furthermore, we propose MusicPro-7k which includes 7,418 film clips, music, description, rhythm spots and main melody, considering the lack of a professional and high-qua

lity film music dataset. Finally, both the standard metrics and the new speciali zed metrics we propose demonstrate that the music generated by our model achieve s state-of-the-art performance in terms of quality, consistency with video, dive rsity, musicality, and musical development. Project page: https://apple-jun.gith ub.io/FilmComposer.github.io/
****

Ground-V: Teaching VLMs to Ground Complex Instructions in Pixels
Yongshuo Zong, Qin Zhang, Dongsheng An, Zhihua Li, Xiang Xu, Linghan Xu, Zhuowen Tu, Yifan Xing, Onkar Dabeer; Proceedings of the Computer Vision and Pattern Re cognition Conference (CVPR), 2025, pp. 24635-24645
This work presents a simple yet effective workflow for automatically scaling ins truction-following data to elicit pixel-level grounding capabilities of VLMs und er complex instructions. In particular, we address five critical real-world chal lenges in text-instruction-based grounding: hallucinated references, multi-objec t scenarios, reasoning, multi-granularity, and part-level references. By leverag ing knowledge distillation from a pre-trained teacher model, our approach genera tes high-quality instruction-response pairs linked to existing pixel-level annot ations, minimizing the need for costly human annotation. The resulting dataset, Ground-V, captures rich object localization knowledge and nuanced pixel-level re ferring expressions. Experiment results show that models trained on Ground-V exh ibit substantial improvements across diverse grounding tasks. Specifically, inco rporating \dataset during training directly achieve an average accuracy boost of 4.4% for LISA and a 7.9% for PSALM across six benchmarks on the gIoU metric. It also sets new state-of-the-art results on standard benchmarks such as RefCOCO/+ /g. Notably, on gRefCOCO, we achieve an N-Acc of 83.3%, exceeding the previous s tate-of-the-art by more than 20%.
****

Structure-from-Motion with a Non-Parametric Camera Model
Yihan Wang, Linfei Pan, Marc Pollefeys, Viktor Larsson; Proceedings of the Compu ter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1040-1049
In this paper, we present a new generic Structure-from-Motion pipeline, GenSfM, that uses a non-parametric camera projection model. The model is self-calibrated during the reconstruction process and can fit a wide variety of cameras, rangin g from simple low-distortion pinhole cameras to more extreme optical systems suc h as fisheye or catadioptric cameras. The key component in our framework is an a daptive calibration procedure that can estimate partial calibrations, only model ing regions of the image where sufficient constraints are available. In experime nts, we show that our method achieves comparable accuracy to traditional Structu re-from-Motion pipelines in easy scenarios, and outperforms them in cases where they are unable to self-calibrate their parametric models. Code is at https://gi thub.com/Ivonne320/GenSfM.git
****

EventPSR: Surface Normal and Reflectance Estimation from Photometric Stereo Usin g an Event Camera
Bohan Yu, Jin Han, Boxin Shi, Imari Sato; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11427-11436
Simultaneously acquisition of the surface normal and reflectance parameters is a crucial but challenging technique in the field of computer vision and graphics. It requires capturing multiple high dynamic range (HDR) images in existing meth ods using frame-based cameras. In this paper, we propose EventPSR, the first wor k to recover surface normal and reflectance parameters (e.g., metallic and rough ness) simultaneously using an event camera. Compared with the existing methods b ased on photometric stereo or neural radiance fields, EventPSR is a robust and e fficient approach that works consistently with different materials. Thanks to th e extremely high temporal resolution and high dynamic range coverage of event ca meras, EventPSR can recover accurate surface normal and reflectance of objects w ith various materials in 10 seconds. Extensive experiments on both synthetic dat a and real objects show that compared with existing methods using more than 100 HDR images, EventPSR recovers comparable surface normal and reflectance paramete rs with only about 30% of the data rate.

****
LAL: Enhancing 3D Human Motion Prediction with Latency-aware Auxiliary Learning
Xiaoning Sun, Dong Wei, Huaijiang Sun, Shengxiang Hu; Proceedings of the Compute r Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7105-7114
Making accurate prediction of human motions based on the historical observation is a crucial technology for robots to collaborate with humans. Existing human mo tion prediction methods are all built under an ideal assumption that robots can instantaneously react, which ignores the time delay introduced during data proce ssing & analysis and future reaction planning -- jointly known as "response late ncy". Consequently, the predictions made within this latency period become meani ngless for practical use, as part of the time has passed and the corresponding r eal motions have already occurred before robot deliver its reaction. In this pap er, we argue that the seemingly meaningless prediction period, however, can be l everaged to enhance prediction accuracy significantly. We propose LAL, a Latency -aware Auxiliary Learning framework, which shifts the existing "reaction instant aneous" convention into a new motion prediction paradigm with both latency compa tibility and utility. The framework consists of two branches handling different tasks: the primary branch learns to directly predict the valid target (excluding the beginning latency period) based on observation; while the auxiliary branch learns the same target, but based on the reformed observation with additional la tency data incorporated. A direct and effective way of auxiliary feature sharing is forced by our tailored consistency loss, to gradually integrate auxiliary la tency insights into the primary prediction branch. Estimated feature statistics- based alignment method is presented as optional step for primary branch refineme nt. Experiments show that LAL achieves significant improvement on prediction acc uracy, without additional time consumption during testing.
****
CASP: Consistency-aware Audio-induced Saliency Prediction Model for Omnidirectio nal Video
Zhaolin Wan, Han Qin, Zhiyang Li, Xiaopeng Fan, Wangmeng Zuo, Debin Zhao; Procee dings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp . 12605-12614
Omnidirectional videos (ODVs) present distinct challenges for accurate audio-vis ual saliency prediction due to their immersive nature, which combines spatial au dio with panoramic visuals to enhance the user experience. While auditory cues a re crucial for guiding visual attention across the panoramic scene, the interact ion between audio and visual stimuli in ODVs remains underexplored. Existing mod els primarily focus on spatiotemporal visual cues and treat audio signals separa tely from their spatial and temporal contexts, often leading to misalignments be tween audio and visual content and undermining temporal consistency across frame s. To bridge these gaps, we propose a novel audio-induced saliency prediction mo del for ODVs that holistically integrates audio and visual inputs through a mult i-modal encoder, an audio-visual interaction module, and an audio-visual transfo rmer. Unlike conventional methods that isolate audio cue locations and attribute s, our model employs a query-based framework, where learnable audio queries capt ure comprehensive audio-visual dependencies, thus enhancing saliency prediction by dynamically aligning with audio cues. Besides, we introduce a novel consisten cy loss to enforce temporal coherence in saliency regions across frames. Extensi ve experiments demonstrate that our model outperforms state-of-the-art methods i n predicting audio-visual salient regions in ODVs, establishing its robustness a nd superior performance.
****
TreeMeshGPT: Artistic Mesh Generation with Autoregressive Tree Sequencing
Stefan Lionar, Jiabin Liang, Gim Hee Lee; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26608-26617
We introduce TreeMeshGPT, an autoregressive Transformer designed to generate hig h-quality artistic meshes aligned with input point clouds. Instead of the conven tional next-token prediction in autoregressive Transformer, we propose a novel A utoregressive Tree Sequencing where the next input token is retrieved from a dyn amically growing tree structure that is built upon the triangle adjacency of fac

es within the mesh. Our sequencing enables the mesh to extend locally from the l ast generated triangular face at each step, and therefore reduces training diffi culty and improves mesh quality. Our approach represents each triangular face wi th two tokens, achieving a compression rate of approximately 22% compared to the naive face tokenization. This efficient tokenization enables our model to gener ate highly detailed artistic meshes with strong point cloud conditioning, surpas sing previous methods in both capacity and fidelity. Furthermore, our method gen erates mesh with strong normal orientation constraints, minimizing flipped norma ls commonly encountered in previous methods. Our experiments show that TreeMeshG PT enhances the mesh generation quality with refined details and normal orientat ion consistency.

****

RefPose: Leveraging Reference Geometric Correspondences for Accurate 6D Pose Est imation of Unseen Objects

Jaeguk Kim, Jaewoo Park, Keuntek Lee, Nam Ik Cho; Proceedings of the Computer Vi sion and Pattern Recognition Conference (CVPR), 2025, pp. 6447-6456

Estimating the 6D pose of unseen objects from monocular RGB images remains a cha llenging problem, especially due to the lack of prior object-specific knowledge. To tackle this issue, we propose RefPose, an innovative approach to object pose estimation that leverages a reference image and geometric correspondence as gui dance. RefPose first predicts an initial pose by using object templates to rende r the reference image and establish the geometric correspondence needed for the refinement stage. During the refinement stage, RefPose estimates the geometric c orrespondence of the query based on the generated references and iteratively ref ines the pose through a render-and-compare approach. To enhance this estimation, we introduce a correlation volume-guided attention mechanism that effectively c aptures correlations between the query and reference images. Unlike traditional methods that depend on pre-defined object models, RefPose dynamically adapts to new object shapes by leveraging a reference image and geometric correspondence. This results in robust performance across previously unseen objects. Extensive e valuation on the BOP benchmark datasets shows that RefPose achieves state-of-the -art results while maintaining a competitive runtime.

****

Relation3D : Enhancing Relation Modeling for Point Cloud Instance Segmentation

Jiahao Lu, Jiacheng Deng; Proceedings of the Computer Vision and Pattern Recogni tion Conference (CVPR), 2025, pp. 8889-8899

3D instance segmentation aims to predict a set of object instances in a scene, r epresenting them as binary foreground masks with corresponding semantic labels. Currently, transformer-based methods are gaining increasing attention due to the ir elegant pipelines and superior predictions. However, these methods primarily focus on modeling the external relationships between scene features and query fe atures through mask attention. They lack effective modeling of the internal rela tionships among scene features as well as between query features.In light of the se disadvantages, we propose Relation3D: Enhancing Relation Modeling for Point C loud Instance Segmentation. Specifically, we introduce an adaptive superpoint ag gregation module and a contrastive learning-guided superpoint refinement module to better represent superpoint features (scene features) and leverage contrastiv e learning to guide the updates of these features.Furthermore, our relation-awar e self-attention mechanism enhances the capabilities of modeling relationships b etween queries by incorporating positional and geometric relationships into the self-attention mechanism.Extensive experiments on the ScanNetV2, ScanNet++, Scan Net200 and S3DIS datasets demonstrate the superior performance of Relation3D.

****

Shape My Moves: Text-Driven Shape-Aware Synthesis of Human Motions

Ting-Hsuan Liao, Yi Zhou, Yu Shen, Chun-Hao Paul Huang, Saayan Mitra, Jia-Bin Hu ang, Uttaran Bhattacharya; Proceedings of the Computer Vision and Pattern Recogn ition Conference (CVPR), 2025, pp. 1917-1928

We explore how body shapes influence human motion synthesis, an aspect often ove rlooked in existing text-to-motion generation methods due to the ease of learnin g a homogenized, canonical body shape. However, this homogenization can distort

the natural correlations between different body shapes and their motion dynamics . Our method addresses this gap by generating body-shape-aware human motions fro m natural language prompts. We utilize a finite scalar quantization-based variat ional autoencoder (FSQ-VAE) to quantize motion into discrete tokens and then lev erage continuous body shape information to de-quantize these tokens back into co ntinuous, detailed motion. Additionally, we harness the capabilities of a pretra ined language model to predict both continuous shape parameters and motion token s, facilitating the synthesis of text-aligned motions and decoding them into sha pe-aware motions. We evaluate our method quantitatively and qualitatively, and a lso conduct a comprehensive perceptual study to demonstrate its efficacy in gene rating shape-aware motions.
****

Generating 3D-Consistent Videos from Unposed Internet Photos
Gene Chou, Kai Zhang, Sai Bi, Hao Tan, Zexiang Xu, Fujun Luan, Bharath Hariharan , Noah Snavely; Proceedings of the Computer Vision and Pattern Recognition Confe rence (CVPR), 2025, pp. 27934-27945
We address the problem of generating videos from unposed internet photos. A hand ful of input images serve as keyframes, and our model interpolates between them to simulate a path moving between the cameras. Given random images, a model's ab ility to capture underlying geometry, recognize scene identity, and relate frame s in terms of camera position and orientation reflects a fundamental understandi ng of 3D structure and scene layout. However, existing video models such as Luma Dream Machine fail at this task. We design a self-supervised method that takes advantage of the consistency of videos and variability of multiview internet pho tos to train a scalable, 3D-aware video model without any 3D annotations such as camera parameters. We validate that our method outperforms commercial models in terms of geometric and appearance consistency. We also show our model benefits applications that enable camera control, such as 3D Gaussian Splatting. Our resu lts suggest that we can scale up scene-level 3D learning using only 2D data such as videos and multiview internet photos.
****

Gazing at Rewards: Eye Movements as a Lens into Human and AI Decision-Making in Hybrid Visual Foraging
Bo Wang, Dingwei Tan, Yen-Ling Kuo, Zhaowei Sun, Jeremy M. Wolfe, Tat-Jen Cham, Mengmi Zhang; Proceedings of the Computer Vision and Pattern Recognition Confere nce (CVPR), 2025, pp. 14810-14823
Imagine searching a collection of coins for quarters (0.25), dimes (0.10), nicke ls (0.05), and pennies (0.01)--a hybrid foraging task where observers search for multiple instances of multiple target types. In such tasks, how do target value s and their prevalence influence foraging and eye movement behaviors (e.g., shou ld you prioritize rare quarters or common nickels)? To explore this, we conducte d human psychophysics experiments, revealing that humans are proficient reward f oragers. Their eye fixations are drawn to regions with higher average rewards, f ixation durations are longer on more valuable targets, and their cumulative rewa rds exceed chance, approaching the upper bound of optimal foragers. To probe the se decision making process of humans, we developed a transformer-based Visual Fo rager (VF) model trained via reinforcement learning. Our VF model takes a series of targets, their corresponding values, and the search image as inputs, process es the images using foveated vision, and produces a sequence of eye movements al ong with decisions on whether to click on each fixated item. Our model outperfor ms all baselines, achieves cumulative rewards comparable to those of humans, and closely mirrors human foraging behavior in eye movements and click biases. Furt hermore, stress tests on out-of-distribution tasks with novel targets, unseen va lues, and varying set sizes demonstrate the VF model's effective generalization. Our work offers valuable insights into the relationship between eye movements a nd decision-making, with our model serving as a powerful tool for further explor ation of this connection. All data, code, and models are available at https://gi thub.com/ZhangLab-DeepNeuroCogLab/visual-forager.
****

FOCUS: Knowledge-enhanced Adaptive Visual Compression for Few-shot Whole Slide I

mage Classification
Zhengrui Guo, Conghao Xiong, Jiabo Ma, Qichen Sun, Lishuang Feng, Jinzhuo Wang, Hao Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15590-15600
Few-shot learning presents a critical solution for cancer diagnosis in computational pathology (CPath), addressing fundamental limitations in data availability, particularly the scarcity of expert annotations and patient privacy constraints. A key challenge in this paradigm stems from the inherent disparity between the limited training set of whole slide images (WSIs) and the enormous number of contained patches, where a significant portion of these patches lacks diagnostically relevant information, potentially diluting the model's ability to learn and focus on critical diagnostic features. While recent works attempt to address this by incorporating additional knowledge, several crucial gaps hinder further progress: (1) despite the emergence of powerful pathology foundation models (FMs), their potential remains largely untapped, with most approaches limiting their use to basic feature extraction; (2) current language guidance mechanisms attempt to align text prompts with vast numbers of WSI patches all at once, struggling to leverage rich pathological semantic information. To this end, we introduce the knowledge-enhanced adaptive visual compression framework, dubbed FOCUS, which uniquely combines pathology FMs with language prior knowledge to enable a focused analysis of diagnostically relevant regions by prioritizing discriminative WSI patches. Our approach implements a progressive three-stage compression strategy: we first leverage FMs for global visual redundancy elimination, and integrate compressed features with language prompts for semantic relevance assessment, then perform neighbor-aware visual token filtering while preserving spatial coherence. Extensive experiments on pathological datasets spanning breast, lung, and ovarian cancers demonstrate its superior performance in few-shot pathology diagnosis. Codes are available at https://github.com/dddavid4real/FOCUS.
****
Beyond Human Perception: Understanding Multi-Object World from Monocular View
Keyu Guo, Yongle Huang, Shijie Sun, Xiangyu Song, Mingtao Feng, Zedong Liu, Huansheng Song, Tiantian Wang, Jianxin Li, Naveed Akhtar, Ajmal Saeed Mian; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3751-3760
Language and binocular vision play a crucial role in human understanding of the world. Advancements in artificial intelligence have also made it possible for machines to develop 3D perception capabilities essential for high-level scene understanding. However, only monocular cameras are often available in practice due to cost and space constraints. Enabling machines to achieve accurate 3D understanding from a monocular view is practical but presents significant challenges. We introduce MonoMulti-3DVG, a novel task aimed at achieving multi-object 3D Visual Grounding (3DVG) based on monocular RGB images, allowing machines to better understand and interact with the 3D world. To this end, we construct a large-scale benchmark dataset, MonoMulti3D-ROPE, and propose a model, CyclopsNet that integrates a State-Prompt Visual Encoder (SPVE) module with a Denoising Alignment Fusion (DAF) module to achieve robust multi-modal semantic alignment and fusion. This leads to more stable and robust multi-modal joint representations for downstream tasks. Experimental results show that our method significantly outperforms existing techniques on the MonoMulti3D-ROPE dataset. Our dataset and code are available at https://github.com/JasonHuang516/MonoMulti-3DVG
****
GRAE-3DMOT: Geometry Relation-Aware Encoder for Online 3D Multi-Object Tracking
Hyunseop Kim, Hyo-Jun Lee, Yonguk Lee, Jinu Lee, Hanul Kim, Yeong Jun Koh; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11697-11706
Recently, 3D multi-object tracking (MOT) has widely adopted the standard tracking-by-detection paradigm, which solves the association problem between detections and tracks. Many tracking-by-detection approaches establish constrained relationships between detections and tracks using a distance threshold to reduce confusion during association. However, this approach does not effectively and comprehe

nsively utilize the information regarding objects due to the constraints of the distance threshold. In this paper, we propose GRAE-3DMOT, Geometry Relation-Aware Encoder 3D Multi-Object Tracking, which contains a geometric relation-aware encoder to produce informative features for association. The geometric relation-aware encoder consists of three components: a spatial relation-aware encoder, a spatiotemporal relation-aware encoder, and a distance-aware feature fusion layer. The spatial relation-aware encoder effectively aggregates detection features by comprehensively exploiting as many detections as possible. The spatiotemporal relation-aware encoder provides spatiotemporal relation-aware features by combing spatial and temporal relation features, where the spatiotemporal relation-aware features are transformed into association scores for MOT. The distance-aware feature fusion layer is integrated into both encoders to enhance the relation features of physically proximate objects. Experimental results demonstrate that the proposed GRAE-3DMOT outperforms the state-of-the-art on the nuScenes. Our approach achieves 73.7% and 70.2% AMOTA on the nuScenes validation and test sets using CenterPoint detections.

****

Automatic Joint Structured Pruning and Quantization for Efficient Neural Network Training and Compression

Xiaoyi Qu, David Aponte, Colby Banbury, Daniel P. Robinson, Tianyu Ding, Kazuhito Koishida, Ilya Zharkov, Tianyi Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15234-15244

Structured pruning and quantization are fundamental techniques used to reduce the size of deep neural networks (DNNs) and typically are applied independently. Applying these techniques jointly via co-optimization has the potential to produce smaller, high-quality models. However, existing joint schemes are not widely used because of (1) engineering difficulties (complicated multi-stage processes), (2) black-box optimization (extensive hyperparameter tuning to control the overall compression), and (3) insufficient architecture generalization. To address these limitations, we present the framework GETA, which automatically and efficiently performs joint structured pruning and quantization-aware training on any DNN. GETA introduces three key innovations: (I) a quantization-aware dependency graph (QADG) that constructs a pruning search space for generic quantization-aware DNN, (II) a partially projected stochastic gradient method that guarantees layer-wise bit constraints are satisfied, and (III) a new joint learning strategy that incorporates interpretable relationships between pruning and quantization. We present numerical experiments on both convolutional neural networks and transformer architectures that show that our approach achieves competitive (often superior) performance compared to existing joint pruning and quantization methods. The source code is available at https://github.com/microsoft/GETA.

****

Parameter Efficient Mamba Tuning via Projector-targeted Diagonal-centric Linear Transformation

Seokil Ham, Hee-Seon Kim, Sangmin Woo, Changick Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30106-30115

Despite the growing interest in Mamba architecture as a potential replacement for Transformer architecture, parameter-efficient fine-tuning (PEFT) approaches for Mamba remain largely unexplored. In our study, we introduce two key insights-driven strategies for PEFT in Mamba architecture: (1) While state-space models (SSMs) have been regarded as the cornerstone of Mamba architecture, then expected to play a primary role in transfer learning, our findings reveal that Projectors ---not SSMs---are the predominant contributors to transfer learning, and (2) Based on our observation, we propose a novel PEFT method specialized to Mamba architecture: Projector-targeted Diagonal-centric Linear Transformation (ProDiaL). ProDiaL focuses on optimizing only diagonal-centric linear transformation matrices, without directly fine-tuning the pretrained Projector weights. This targeted approach allows efficient task adaptation, utilizing less than 1% of the total parameters, and exhibits strong performance across both vision and language Mamba models, highlighting its versatility and effectiveness.

****

ViUniT: Visual Unit Tests for More Robust Visual Programming
Artemis Panagopoulou, Honglu Zhou, Silvio Savarese, Caiming Xiong, Chris Callison-Burch, Mark Yatskar, Juan Carlos Niebles; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24646-24656
Programming based approaches to reasoning tasks have substantially expanded the types of questions models can answer about visual scenes.Yet on benchmark visual reasoning data, when answering correctly, such models produce incorrect programs 33% of the time. These models are often right for the wrong reasons and risk unexpected failures on new data. Unit tests play a foundational role in ensuring code correctness and could be used to repair such failures. We propose Visual Unit Testing (ViUniT), a framework to improve the reliability of visual programs by automatically generating unit tests. In our framework, a unit test is represented as a novel image and answer meant to verify logical correctness of a program produced for a given query. Our method leverages a language model to create unit tests in the form of image descriptions and expected answers and image synthesis to produce corresponding images. We conduct a comprehensive analysis of what constitutes an effective visual unit test suite, exploring unit test generation, sampling strategies, image generation methods, and varying the number of programs and unit tests. Additionally, we introduce four applications of visual unit tests: best program selection, answer refusal, re-prompting, and unsupervised reward formulations for reinforcement learning. Experiments with two models across three datasets in visual question answering and image-text matching demonstrate that ViUniT improves model performance by 11.4%. Notably, it enables 7B open-source models to outperform gpt-4o-mini by an average of 7.7% and reduces the occurrence of programs that are correct for the wrong reasons by 40%.
****
LIRM: Large Inverse Rendering Model for Progressive Reconstruction of Shape, Materials and View-dependent Radiance Fields
Zhengqin Li, Dilin Wang, Ka Chen, Zhaoyang Lv, Thu Nguyen-Phuoc, Milim Lee, Jia-Bin Huang, Lei Xiao, Yufeng Zhu, Carl S. Marshall, Yuheng Ren, Richard Newcombe, Zhao Dong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 505-517
We present Large Inverse Rendering Model (LIRM), a transformer architecture that jointly reconstructs high-quality shape, materials, and radiance fields with view-dependent effects in less than a second. Our model builds upon the recent Large Reconstruction Models (LRMs) that achieve state-of-the-art sparse-view reconstruction quality. However, existing LRMs struggle to reconstruct unseen parts accurately and cannot recover glossy appearance or generate relightable 3D contents that can be consumed by standard Graphics engines. To address these limitations, we make three key technical contributions to build a more practical multi-view 3D reconstruction framework. First, we introduce an update model that allows us to progressively add more input views to improve our reconstruction. Second, we propose a hexa-plane neural SDF representation to better recover detailed textures, geometry and material parameters. Third, we develop a novel neural directional-embedding mechanism to handle view-dependent effects. Trained on a large-scale shape and material dataset with a tailored coarse-to-fine training scheme, our model achieves compelling results. It compared favorably to optimization-based dense-view inverse rendering methods in terms of geometry and relighting accuracy, while requiring only a fraction of the inference time.
****
DualTalk: Dual-Speaker Interaction for 3D Talking Head Conversations
Ziqiao Peng, Yanbo Fan, Haoyu Wu, Xuan Wang, Hongyan Liu, Jun He, Zhaoxin Fan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21055-21064
In face-to-face conversations, individuals need to switch between speaking and listening roles seamlessly. Existing 3D talking head generation models focus solely on speaking or listening, neglecting the natural dynamics of interactive conversation, which leads to unnatural interactions and awkward transitions. To address this issue, we propose a new task--multi-round dual-speaker interaction for 3D talking head generation--which requires models to handle and generate both sp

eaking and listening behaviors in continuous conversation. To solve this task, w
e introduce DualTalk, a novel unified framework that integrates the dynamic beha
viors of speakers and listeners to simulate realistic and coherent dialogue inte
ractions. This framework not only synthesizes lifelike talking heads when speaki
ng but also generates continuous and vivid non-verbal feedback when listening, e
ffectively capturing the interplay between the roles. We also create a new datas
et featuring 50 hours of multi-round conversations with over 1,000 characters, w
here participants continuously switch between speaking and listening roles. Exte
nsive experiments demonstrate that our method significantly enhances the natural
ness and expressiveness of 3D talking heads in dual-speaker conversations. We re
commend watching the supplementary video: https://ziqiaopeng.github.io/dualtalk
****

MAR-3D: Progressive Masked Auto-regressor for High-Resolution 3D Generation
Jinnan Chen, Lingting Zhu, Zeyu Hu, Shengju Qian, Yugang Chen, Xin Wang, Gim Hee
 Lee; Proceedings of the Computer Vision and Pattern Recognition Conference (CVP
R), 2025, pp. 11083-11092
Recent advances in auto-regressive transformers have revolutionized generative m
odeling across different domains, from language processing to visual generation,
 demonstrating remarkable capabilities. However, applying these advances to 3D g
eneration presents three key challenges: the unordered nature of 3D data conflic
ts with sequential next-token prediction paradigm, conventional vector quantizat
ion approaches incur substantial compression loss when applied to 3D meshes, and
 the lack of efficient scaling strategies for higher resolution latent predictio
n. To address these challenges, we introduce MAR-3D, which integrates a pyramid
variational autoencoder with a cascaded masked auto-regressive transformer (Casc
aded MAR) for progressive latent upscaling in the continuous space. Our architec
ture employs random masking during training and auto-regressive denoising in ran
dom order during inference, naturally accommodating the unordered property of 3D
 latent tokens. Additionally, we propose a cascaded training strategy with condi
tion augmentation that enables efficiently up-scale the latent token resolution
with fast convergence. Extensive experiments demonstrate that MAR-3D not only ac
hieves superior performance and generalization capabilities compared to existing
 methods but also exhibits enhanced scaling capabilities compared to joint distr
ibution modeling approaches (e.g., diffusion transformers).
****

beta-FFT: Nonlinear Interpolation and Differentiated Training Strategies for Sem
i-Supervised Medical Image Segmentation
Ming Hu, Jianfu Yin, Zhuangzhuang Ma, Jianheng Ma, Feiyu Zhu, Bingbing Wu, Ya We
n, Meng Wu, Cong Hu, Bingliang Hu, Quan Wang; Proceedings of the Computer Vision
 and Pattern Recognition Conference (CVPR), 2025, pp. 30839-30849
Co-training has achieved significant success in the field of semi-supervised lea
rning; however, the *homogenization phenomenon*, which arises from multiple mode
ls tending towards similar decision boundaries, remains inadequately addressed.
To tackle this issue, we propose a novel algorithm called **β-FFT** from th
e perspectives of data diversity and training structure.First, from the perspect
ive of data diversity, we introduce a nonlinear interpolation method based on th
e **Fast Fourier Transform (FFT)**. This method generates more diverse training
samples by swapping low-frequency components between pairs of images, thereby en
hancing the model's generalization capability. Second, from the structural persp
ective, we propose a differentiated training strategy to alleviate the homogeniz
ation issue in co-training. In this strategy, we apply additional training with
labeled data to one model in the co-training framework, while employing linear i
nterpolation based on the **Beta (β)** distribution for the unlabeled data
as a regularization term for the additional training. This approach enables us t
o effectively utilize the limited labeled data while simultaneously improving th
e model's performance on unlabeled data, ultimately enhancing the overall perfor
mance of the system.Experimental results demonstrate that **β-FFT** outperf
orms current state-of-the-art (SOTA) methods on three public medical image datas
ets.
****

Dynamic Group Normalization: Spatio-Temporal Adaptation to Evolving Data Statistics

Yair Smadar, Assaf Hoogi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30167-30177

Deep neural networks remain vulnerable to statistical variations in data, even with advances in normalization techniques. Existing methods use fixed-size normalization sets, restricting their adaptability to evolving and diverse data characteristics. We introduce Dynamic Group Normalization (DGN), a framework that dynamically adjusts channel grouping based on statistical awareness, ensuring a flexible and optimized formation of channel groups. By leveraging an efficient spatio-temporal mechanism, DGN continuously evaluates inter-channel relationships within layers and across training epochs, ensuring robust and responsive adaptation to the updated data statistics. Extensive evaluations of 34 architectures and 11 computer vision benchmarks show the consistent superiority of DGN over traditional normalization methods. It achieves significant accuracy gains in classification, detection, and segmentation tasks while maintaining computational efficiency. Moreover, it outperforms traditional methods in challenging scenarios, including out-of-distribution generalization, imbalanced long-tailed distributions, and corrupted data.

****

Latent Space Super-Resolution for Higher-Resolution Image Generation with Diffusion Models

Jinho Jeong, Sangmin Han, Jinwoo Kim, Seon Joo Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2355-2365

In this paper, we propose LSRNA, a novel framework for higher-resolution (exceeding 1K) image generation using diffusion models by leveraging super-resolution directly in the latent space. Existing diffusion models struggle with scaling beyond their training resolutions, often leading to structural distortions or content repetition. Reference-based methods address the issues by upsampling a low-resolution reference to guide higher-resolution generation. However, they face significant challenges: upsampling in latent space often causes manifold deviation, which degrades output quality. On the other hand, upsampling in RGB space tends to produce overly smoothed outputs. To overcome these limitations, LSRNA combines Latent space Super-Resolution (LSR) for manifold alignment and Region-wise Noise Addition (RNA) to enhance high-frequency details. Our extensive experiments demonstrate that integrating LSRNA outperforms state-of-the-art reference-based methods across various resolutions and metrics, while showing the critical role of latent space upsampling in preserving detail and sharpness. The code will be available at https://github.com/3587jjh/LSRNA https://github.com/3587jjh/LSRNA.

****

SynerGen-VL: Towards Synergistic Image Understanding and Generation with Vision Experts and Token Folding

Hao Li, Changyao Tian, Jie Shao, Xizhou Zhu, Zhaokai Wang, Jinguo Zhu, Wenhan Dou, Xiaogang Wang, Hongsheng Li, Lewei Lu, Jifeng Dai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29767-29779

The remarkable success of Large Language Models (LLMs) has extended to the multimodal domain, achieving outstanding performance in image understanding and generation. Recent efforts to develop unified Multimodal Large Language Models (MLLMs) that integrate these capabilities have shown promising results. However, existing approaches often involve complex designs in model architecture or training pipeline, increasing the difficulty of model training and scaling. In this paper, we propose SynerGen-VL, a simple yet powerful encoder-free MLLM capable of both image understanding and generation. To address challenges identified in existing encoder-free unified MLLMs, we introduce the token folding mechanism and the vision-expert-based progressive alignment pretraining strategy, effectively supporting high-resolution image understanding while reducing training complexity. After being trained on large-scale mixed image-text data with a unified next-token prediction objective, SynerGen-VL achieves or surpasses the performance of existing encoder-free unified MLLMs with comparable or smaller parameter sizes, and narrows the gap with task-specific state-of-the-art models, highlighting a promi

sing path toward future unified MLLMs. Our code and models shall be released.
****
Synthetic Prior for Few-Shot Drivable Head Avatar Inversion
Wojciech Zielonka, Stephan J. Garbin, Alexandros Lattas, George Kopanas, Paulo G otardo, Thabo Beeler, Justus Thies, Timo Bolkart; Proceedings of the Computer Vi sion and Pattern Recognition Conference (CVPR), 2025, pp. 10735-10746
We present SynShot, a novel method for the few-shot inversion of a drivable head avatar based on a synthetic prior. We tackle three major challenges. First, tra ining a controllable 3D generative network requires a large number of diverse se quences, for which pairs of images and high-quality tracked meshes are not alway s available. Second, the use of real data is strictly regulated (e.g., under the General Data Protection Regulation, which mandates frequent deletion of models and data to accommodate a situation when participant's consent is withdrawn). Sy nthetic data, free from these constraints, is an appealing alternative. Third, s tate-of-the-art monocular avatar models struggle to generalize to new views and expressions, lacking a strong prior and often overfitting to a specific viewpoin t distribution. Inspired by machine learning models trained solely on synthetic data, we propose a method that learns a prior model from a large dataset of synt hetic heads with diverse identities, expressions, and viewpoints. With few input images, SynShot fine-tunes the pretrained synthetic prior to bridge the domain gap, modeling a photorealistic head avatar that generalizes to novel expressions and viewpoints. We model the head avatar using 3D Gaussian splatting and a conv olutional encoder-decoder that outputs Gaussian parameters in UV texture space. To account for the different modeling complexities over parts of the head (e.g., skin vs hair), we embed the prior with explicit control for upsampling the numb er of per-part primitives. Compared to SOTA monocular and GAN-based methods, Syn Shot significantly improves novel view and expression synthesis.
****
Reasoning in Visual Navigation of End-to-end Trained Agents: A Dynamical Systems Approach
Steeven Janny, Hervé Poirier, Leonid Antsfeld, Guillaume Bono, Gianluca Monaci, Boris Chidlovskii, Francesco Giuliari, Alessio Del Bue, Christian Wolf; Proceedi ngs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12111-12121
Progress in Embodied AI has made it possible for end-to-end-trained agents to na vigate in photo-realistic environments with high-level reasoning and zero-shot o r language-conditioned behavior, but evaluations and benchmarks are still domina ted by simulation. In this work, we focus on the fine-grained behavior of fast-m oving real robots and present a large-scale experimental study involving \numepi sodes navigation episodes in a real environment with a physical robot, where w e analyze the type of reasoning emerging from end-to-end training. In particular , we study the presence of realistic dynamics which the agent learned for open-l oop forecasting, and their interplay with sensing. We analyze the way the agent uses latent memory to hold elements of the scene structure and information gathe red during exploration. We probe the planning capabilities of the agent, and fin d in its memory evidence for somewhat precise plans over a limited horizon. Furt hermore, we show in a post-hoc analysis that the value function learned by the a gent relates to long-term planning. Put together, our experiments paint a new pi cture on how using tools from computer vision and sequential decision making hav e led to new capabilities in robotics and control. An interactive tool is availa ble at https://visual-navigation-reasoning.github.io
****
Rethinking Noisy Video-Text Retrieval via Relation-aware Alignment
Huakai Lai, Guoxin Xiong, Huayu Mai, Xiang Liu, Tianzhu Zhang; Proceedings of th e Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9231-9241
Video-Text Retrieval (VTR) is a core task in multi-modal understanding, drawing growing attention from both academia and industry in recent years. While numerou s VTR methods have achieved success, most of them assume accurate visual-text co rrespondences during training, which is difficult to ensure in practice due to u biquitous noise, known as noisy correspondences (NC). In this paper, we rethink

how to mitigate the NC from the perspective of representative reference features (termed agents), and propose a novel relation-aware purified consistency (RPC) network to amend direct pairwise correlation, including representative agents construction and relation-aware ranking distribution alignment. The proposed RPC enjoys several merits. First, to learn the agents well without any correspondence supervision, we customize the agents construction according to the three characteristics of reliability, representativeness, and resilience. Second, the ranking distribution-based alignment process leverages the structural information inherent in inter-pair relationships, making it more robust compared to individual comparisons. Extensive experiments on five datasets under different settings demonstrate the efficacy and robustness of our method.
****

DFormerv2: Geometry Self-Attention for RGBD Semantic Segmentation
Bo-Wen Yin, Jiao-Long Cao, Ming-Ming Cheng, Qibin Hou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19345-19355
Recent advances in scene understanding benefit a lot from depth maps because of the 3D geometry information, especially in complex conditions (e.g., low light and overexposed). Existing approaches encode depth maps along with RGB images and perform feature fusion between them to enable more robust predictions. Taking into account that depth can be regarded as a geometry supplement for RGB images, a straightforward question arises: Do we really need to explicitly encode depth information with neural networks as done for RGB images? Based on this insight, in this paper, we investigate a new way to learn RGBD feature representations and present DFormerv2, a strong RGBD encoder that explicitly uses depth maps as geometry priors rather than encoding depth information with neural networks. Our goal is to extract the geometry clues from the depth and spatial distances among all the image patch tokens, which will then be used as geometry priors to allocate attention weights in self-attention. Extensive experiments demonstrate that DFormerv2 exhibits exceptional performance in various RGBD semantic segmentation benchmarks. Code is available at: https://github.com/VCIP-RGBD/DFormer.
****

Scaling Vision Pre-Training to 4K Resolution
Baifeng Shi, Boyi Li, Han Cai, Yao Lu, Sifei Liu, Marco Pavone, Jan Kautz, Song Han, Trevor Darrell, Pavlo Molchanov, Hongxu Yin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9631-9640
High-resolution perception of visual details is crucial for daily tasks. Current vision pre-training, however, is still limited to low resolutions (e.g., 378 x 378 pixels) due to the quadratic cost of processing larger images. We introduce PS3 that scales CLIP-style vision pre-training to 4K resolution with a near-constant cost. Instead of contrastive learning on global image representation, PS3 is pre-trained by selectively processing local regions and contrasting them with local detailed captions, enabling high-resolution representation learning with greatly reduced computational overhead. The pre-trained PS3 is able to both encode the global image at low resolution and selectively process local high-resolution regions based on their saliency or relevance to a text prompt. When applying PS3 to multi-modal LLM (MLLM), the resulting model, named VILA-HD, significantly improves high-resolution visual perception compared to baselines without high-resolution vision pre-training such as AnyRes and S^2 while using up to 4.3x fewer tokens. PS3 also unlocks appealing scaling properties of VILA-HD, including scaling up resolution for free and scaling up test-time compute for better performance. Compared to state of the arts, VILA-HD outperforms previous MLLMs such as NVILA and Qwen2-VL across multiple benchmarks and achieves better efficiency than latest token pruning approaches. Finally, we find current benchmarks do not require 4K-resolution perception, which motivates us to propose 4KPro, a new benchmark of image QA at 4K resolution, on which VILA-HD outperforms all previous MLLMs, including a 14.5% improvement over GPT-4o, and a 3.2% improvement and 2.96x speedup over Qwen2-VL.
****

GarmentPile: Point-Level Visual Affordance Guided Retrieval and Adaptation for Cluttered Garments Manipulation

Ruihai Wu, Ziyu Zhu, Yuran Wang, Yue Chen, Jiarui Wang, Hao Dong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6950-6959

Cluttered garments manipulation poses significant challenges in robotics due to the complex, deformable nature of garments and intricate garment relations. Unlike single-garment manipulation, cluttered scenarios require managing complex garment entanglements and interactions, while maintaining garment cleanliness and manipulation stability. To address these demands, we propose to learn point-level affordance, the dense representation modeling the complex space and multi-modal manipulation candidates, with novel designs for the awareness of garment geometry, structure, and inter-object relations. Additionally, we introduce an adaptation module, informed by learned affordance, to reorganize cluttered garments into configurations conducive to manipulation. Our framework demonstrates effectiveness over environments featuring diverse garment types and pile scenarios in both simulation and the real world. Project page: https://garmentpile.github.io/.

****

Uncertain Multimodal Intention and Emotion Understanding in the Wild

Qu Yang, Qinghongya Shi, Tongxin Wang, Mang Ye; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24700-24709

Understanding intention and emotion from social media poses unique challenges due to the inherent uncertainty in multimodal data, where posts often contain incomplete or missing modalities. While this uncertainty reflects real-world scenarios, it remains underexplored within the computer vision community, particularly in conjunction with the intrinsic relationship between emotion and intention. To address these challenges, we introduce the Multimodal IntentioN and Emotion Understanding in the Wild (MINE) dataset, comprising over 20,000 topic-specific social media posts with natural modality variations across text, image, video, and audio. MINE is distinctively constructed to capture both the uncertain nature of multimodal data and the implicit correlations between intentions and emotions, providing extensive annotations for both aspects. To tackle these scenarios, we propose the Bridging Emotion-Intention via Implicit Label Reasoning (BEAR) framework. BEAR consists of two key components: a BEIFormer that leverages emotion-intention correlations, and a Modality Asynchronous Prompt that handles modality uncertainty. Experiments show that BEAR outperforms existing methods in processing uncertain multimodal data while effectively mining emotion-intention relationships for social media content understanding.

****

GroomLight: Hybrid Inverse Rendering for Relightable Human Hair Appearance Modeling

Yang Zheng, Menglei Chai, Delio Vicini, Yuxiao Zhou, Yinghao Xu, Leonidas Guibas, Gordon Wetzstein, Thabo Beeler; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16040-16050

We present GroomLight, a novel method for relightable hair appearance modeling from multi-view images. Existing hair capture methods struggle to balance photorealistic rendering with relighting capabilities. Analytical material models, while physically grounded, often fail to fully capture appearance details. Conversely, neural rendering approaches excel at view synthesis but generalize poorly to novel lighting conditions. GroomLight addresses this challenge by combining the strengths of both paradigms. It employs an extended hair BSDF model to capture primary light transport and a light-aware residual model to reconstruct the remaining details. We further propose a hybrid inverse rendering pipeline to optimize both components, enabling high-fidelity relighting, view synthesis, and material editing. Extensive evaluations on real-world hair data demonstrate state-of-the-art performance of our method. Our project website is at: https://syntec-research.github.io/GroomLight.

****

Improving Editability in Image Generation with Layer-wise Memory

Daneul Kim, Jaeah Lee, Jaesik Park; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7889-7898

Most real-world image editing tasks require multiple sequential edits to achieve

desired results. Current editing approaches, primarily designed for single-object modifications, struggle with sequential editing: especially with maintaining previous edits along with adapting new objects naturally into the existing content. These limitations significantly hinder complex editing scenarios where multiple objects need to be modified while preserving their contextual relationships. We address this fundamental challenge through two key proposals: enabling rough mask inputs that preserve existing content while naturally integrating new elements and supporting consistent editing across multiple modifications. Our framework achieves this through layer-wise memory, which stores latent representations and prompt embeddings from previous edits. We propose Background Consistency Guidance that leverages memorized latents to maintain scene coherence and Multi-Query Disentanglement in cross-attention that ensures natural adaptation to existing content. To evaluate our method, we present a new benchmark dataset incorporating semantic alignment metrics and interactive editing scenarios. Through comprehensive experiments, we demonstrate superior performance in iterative image editing tasks with minimal user effort, requiring only rough masks while maintaining high-quality results throughout multiple editing steps.

****

Sea-ing in Low-light

Nisha Varghese, A. N. Rajagopalan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16629-16640

Underwater (UW) robotics applications require depth and restored images simultaneously in real-time, irrespective of whether the UW images are captured in good lighting conditions or not. Most of the UW image restoration and depth estimation methods have been devised for images under normal lighting. Consequently, they struggle to perform on poorly lit images. Even though artificial illumination can be used when there is insufficient ambient light, it can introduce non-uniform lighting artifacts in the restored images. Hence, the recovery of depth and restored images directly from Low-Light UW (LLUW) images is a critical requirement in marine applications. While a few works have attempted LLUW image restoration, there are no reported works on joint recovery of depth and clean image from LLUW images. We propose a Self-supervised Low-light Underwater Image and Depth recovery network (SelfLUID-Net) for joint estimation of depth and restored image in real-time from a single LLUW image. We have collected an Underwater Low-light Stereo Video (ULVStereo) dataset which is the first-ever UW dataset with stereo pairs of low-light and normally-lit UW images. For the dual tasks of image and depth recovery from a LLUW image, we effectively utilize the stereo data from ULVStereo that provides cues for both depth and illumination-independent clean image. We harness a combination of the UW image formation process, the Retinex model, and constraints enforced by the scene geometry for our self-supervised training. To handle occlusions, we additionally utilize monocular frames from our video dataset and propose a masking scheme to prevent dynamic transients, that do not respect the underlying scene geometry, from misguiding the learning process. Evaluations on five LLUW datasets demonstrate the superiority and generalization ability of our proposed SelfLUID-Net over existing state-of-the-art methods. The dataset ULVStereo is available at https://github.com/nishavarghese15/ULVStereo.

****

VidTwin: Video VAE with Decoupled Structure and Dynamics

Yuchi Wang, Junliang Guo, Xinyi Xie, Tianyu He, Xu Sun, Jiang Bian; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22922-22932

Recent advancements in video autoencoders (Video AEs) have significantly improved the quality and efficiency of video generation. In this paper, we propose a novel and compact video autoencoder, VidTwin, that decouples video into two distinct latent spaces: Structure latent vectors, which capture overall content and global movement, and Dynamics latent vectors, which represent fine-grained details and rapid movements. Specifically, our approach leverages an Encoder-Decoder backbone, augmented with two submodules for extracting these latent spaces, respectively. The first submodule employs a Q-Former to extract low-frequency motion trends, followed by downsampling blocks to remove redundant content details. The

second averages the latent vectors along the spatial dimension to capture rapid motion. Extensive experiments show that VidTwin achieves a high compression rate of 0.2% with high reconstruction quality (PSNR of 28.14 on the MCL-JCV dataset), and performs efficiently and effectively in downstream generative tasks. Moreover, our model demonstrates explainability and scalability, paving the way for future research in video latent representation and generation.

****

CL-LoRA: Continual Low-Rank Adaptation for Rehearsal-Free Class-Incremental Learning

Jiangpeng He, Zhihao Duan, Fengqing Zhu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30534-30544

Class-Incremental Learning (CIL) aims to learn new classes sequentially while retaining the knowledge of previously learned classes. Recently, pre-trained models (PTMs) combined with parameter-efficient fine-tuning (PEFT) have shown remarkable performance in rehearsal-free CIL without requiring exemplars from previous tasks. However, existing adapter-based methods, which incorporate lightweight learnable modules into PTMs for CIL, create new adapters for each new task, leading to both parameter redundancy and failure to leverage shared knowledge across tasks. In this work, we propose ContinuaL Low-Rank Adaptation (CL-LoRA), which introduces a novel dual-adapter architecture combining task-shared adapters to learn cross-task knowledge and task-specific adapters to capture unique features of each new task. Specifically, the shared adapters utilize random orthogonal matrices and leverage knowledge distillation with gradient reassignment to preserve essential shared knowledge. In addition, we introduce learnable block-wise weights for task-specific adapters, which mitigate inter-task interference while maintaining the model's plasticity. We demonstrate CL-LoRA consistently achieves promising performance under multiple benchmarks with reduced training and inference computation, establishing a more efficient and scalable paradigm for continual learning with pre-trained models.

****

Generative Modeling of Class Probability for Multi-Modal Representation Learning

JungKyoo Shin, Bumsoo Kim, Eunwoo Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20737-20746

Multi-modal understanding plays a crucial role in artificial intelligence by enabling models to jointly interpret inputs from different modalities. However, conventional approaches such as contrastive learning often struggle with modality discrepancies, leading to potential misalignments. In this paper, we propose a novel class anchor alignment approach that leverages class probability distributions for multi-modal representation learning. Our method, Class-anchor-ALigned generative Modeling (CALM), encodes class anchors as prompts to generate and align class probability distributions for each modality, enabling more flexible alignment. Furthermore, we introduce a cross-modal probabilistic variational autoencoder to model uncertainty in the alignment, enhancing the ability to capture deeper relationships between modalities and data variations. Extensive experiments on four benchmark datasets demonstrate that our approach significantly outperforms state-of-the-art methods, especially in out-of-domain evaluations. This highlights its superior generalization capabilities in multi-modal representation learning.

****

VisionZip: Longer is Better but Not Necessary in Vision Language Models

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, Jiaya Jia; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19792-19802

Recent advancements in vision-language models have enhanced performance by increasing the length of visual tokens, making them much longer than text tokens and significantly raising computational costs.However, we observe that the visual tokens generated by popular vision encoders, such as CLIP and SigLIP, contain significant redundancy. To address this, we introduce VisionZip, a simple yet effective method that selects a set of informative tokens for input to the language model, reducing visual token redundancy and improving efficiency while maintaining

model performance. The proposed VisionZip can be widely applied to image and video understanding tasks and is well-suited for multi-turn dialogues in real-world scenarios, where previous methods tend to underperform.Experimental results show that VisionZip outperforms the previous state-of-the-art method by at least 5% performance gains across nearly all settings.Moreover, our method significantly enhances model inference speed, improving the prefilling time by 8xand enabling the LLaVA-Next 13B model to infer faster than the LLaVA-Next 7B model while achieving better results.Furthermore, we analyze the causes of this redundancy and encourage the community to focus on extracting better visual features rather than merely increasing token length. Our code is available at https://github.com/dvlab-research/VisionZip.

****

Simplification Is All You Need against Out-of-Distribution Overconfidence
Keke Tang, Chao Hou, Weilong Peng, Xiang Fang, Zhize Wu, Yongwei Nie, Wenping Wang, Zhihong Tian; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5030-5040
Deep neural networks (DNNs) often exhibit out-of-distribution (OOD) overconfidence, producing overly confident predictions on OOD samples. We attribute this issue to the inherent over-complexity of DNNs and investigate two key aspects: capacity and nonlinearity. First, we demonstrate that reducing model capacity through knowledge distillation can effectively mitigate OOD overconfidence. Second, we show that selectively reducing nonlinearity by removing ReLU operations further alleviates the issue. Building on these findings, we present a practical guide to model simplification, combining both strategies to significantly reduce OOD overconfidence. Extensive experiments validate the effectiveness of this approach in mitigating OOD overconfidence and demonstrate its superiority over state-of-the-art methods. Additionally, our simplification strategies can be combined with existing OOD detection techniques to further enhance OOD detection performance.

****

SpatialDreamer: Self-supervised Stereo Video Synthesis from Monocular Input
Zhen Lv, Yangqi Long, Congzhentao Huang, Cao Li, Chengfei Lv, Hao Ren, Dian Zheng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 811-821
Stereo video synthesis from monocular input is challenging in spatial computing and virtual reality due to the lack of high-quality stereo video pairs for training and the difficulty of maintaining spatio-temporal consistency between frames. Existing methods primarily address these issues by directly applying novel view synthesis (NVS) techniques to video, while facing limitations such as the inability to effectively represent dynamic scenes and the requirement for extensive training data. In this paper, we introduce a novel self-supervised stereo video synthesis paradigm via a video diffusion model, termed SpatialDreamer, which meets the challenges head-on. Firstly, to address the stereo video data insufficiency, we propose a Depth based Video Generation module DVG, which employs a forward-backward rendering mechanism to generate paired videos with geometric and temporal priors. Leveraging data generated by DVG, we propose RefinerNet along with a self-supervised synthetic framework designed to facilitate efficient and dedicated training.More importantly, we devise a consistency control module, which consists of a metric of stereo deviation strength and a Temporal Interaction Learning module TIL for geometric and temporal consistency ensurance respectively. We evaluated the proposed method against various benchmark methods, with the results showcasing its superior performance.

****

LOD-GS: Achieving Levels of Detail using Scalable Gaussian Soup
Jianxiong Shen, Yue Qian, Xiaohang Zhan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 671-680
Current 3D Gaussian Splatting methods often overlook structured information, leading to disorganized 3D Gaussians that compromise memory and structure efficiency, particularly in Level-of-Detail (LOD) management. This inefficiency results in excessive memory use, limiting their application in resource-sensitive environ

ments like virtual reality and interactive simulations. To overcome these challe nges, we introduce a scalable Gaussian Soup that enables high-performance LOD ma nagement with progressively reduced memory usage. Our method utilizes triangle p rimitives with Gaussian splats embedded and adaptive pruning/growing strategies to ensure high-quality scene rendering with significantly reduced memory demands . By embedding neural Gaussians within the triangle primitives through the trian gle-MLP, we achieve further memory savings while maintaining rendering fidelity. Experimental results demonstrate that our approach achieves consistently superi or performance than recent leading techniques across various LOD while progressi vely reducing memory usage.
****

BlenderGym: Benchmarking Foundational Model Systems for Graphics Editing
Yunqi Gu, Ian Huang, Jihyeon Je, Guandao Yang, Leonidas Guibas; Proceedings of t he Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18574-18 583
3D graphics editing is crucial in applications like movie production and game de sign, yet it remains a time-consuming process that demands highly specialized do main expertise. Automating this process is challenging because graphical editing requires performing a variety of tasks, each requiring distinct skill sets. Rec ently, vision-language models (VLMs) have emerged as a powerful framework for au tomating the editing process, but their development and evaluation are bottlenec ked by the lack of a comprehensive benchmark that requires human-level perceptio n and presents real-world editing complexity. In this work, we present BlenderGy m, the first comprehensive VLM system benchmark for 3D graphics editing. Blender Gym evaluates VLM systems through code-based 3D reconstruction tasks. We evaluat e closed- and open-source VLM systems and observe that even the state-of-the-art VLM system struggles with tasks relatively easy for human Blender users. Enable d by BlenderGym, we study how inference scaling techniques impact VLM's performa nce on graphics editing tasks. Notably, our findings reveal that the verifier us ed to guide the scaling of generation can itself be improved through inference s caling, complementing recent insights on inference scaling of LLM generation in coding and math tasks. We further show that inference compute is not uniformly e ffective and can be optimized by strategically distributing it between generatio n and verification.
****

VoteFlow: Enforcing Local Rigidity in Self-Supervised Scene Flow
Yancong Lin, Shiming Wang, Liangliang Nan, Julian Kooij, Holger Caesar; Proceedi ngs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17155-17164
Scene flow estimation aims to recover per-point motion from two adjacent LiDAR s cans. However, in real-world applications such as autonomous driving, points rar ely move independently of others, especially for nearby points belonging to the same object, which often share the same motion. Incorporating this locally rigid motion constraint has been a key challenge in self-supervised scene flow estima tion, which is often addressed by post-processing or appending extra regularizat ion. While these approaches are able to improve the rigidity of predicted flows, they lack an architectural inductive bias for local rigidity within the model s tructure, leading to suboptimal learning efficiency and inferior performance. In contrast, we enforce local rigidity with a lightweight add-on module in neural network design, enabling end-to-end learning. We design a discretized voting spa ce that accommodates all possible translations and then identify the one shared by nearby points by differentiable voting. Additionally, to ensure computationa l efficiency, we operate on pillars rather than points and learn representative features for voting per pillar. We plug the Voting Module into popular model des igns and evaluate its benefit on Argoverse 2 and Waymo datasets. We outperform b aseline works with only marginal compute overhead. Code is available at https:// github.com/tudelft-iv/VoteFlow.
****

The Devil is in Low-Level Features for Cross-Domain Few-Shot Segmentation
Yuhan Liu, Yixiong Zou, Yuhua Li, Ruixuan Li; Proceedings of the Computer Vision

and Pattern Recognition Conference (CVPR), 2025, pp. 4618-4627
Cross-Domain Few-Shot Segmentation (CDFSS) is proposed to transfer the pixel-level segmentation capabilities learned from large-scale source-domain datasets to downstream target-domain datasets, with only a few annotated images per class. In this paper, we focus on a well-observed but unresolved phenomenon in CDFSS: for target domains, particularly those distant from the source domain, segmentation performance peaks at the very early epochs, and declines sharply as the source-domain training proceeds. We delve into this phenomenon for an interpretation: low-level features are vulnerable to domain shifts, leading to sharper loss landscapes during the source-domain training, which is the devil of CDFSS. Based on this phenomenon and interpretation, we further propose a method that includes two plug-and-play modules: one to flatten the loss landscapes for low-level features during source-domain training as a novel sharpness-aware minimization method, and the other to directly supplement target-domain information to the model during target-domain testing by low-level-based calibration. Extensive experiments on four target datasets validate our rationale and demonstrate that our method surpasses the state-of-the-art method in CDFSS signifcantly by 3.71% and 5.34% average MIoU in 1-shot and 5-shot scenarios, respectively.
****

Design2GarmentCode: Turning Design Concepts to Tangible Garments Through Program Synthesis
Feng Zhou, Ruiyang Liu, Chen Liu, Gaofeng He, Yong-Lu Li, Xiaogang Jin, Huamin Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23712-23722
Sewing patterns, the essential blueprints for fabric cutting and tailoring, act as a crucial bridge between design concepts and producible garments. However, existing uni-modal sewing pattern generation models struggle to effectively encode complex design concepts with a multi-modal nature and correlate them with vectorized sewing patterns that possess precise geometric structures and intricate sewing relations. In this work, we propose a novel sewing pattern generation approach Design2GarmentCode based on Large Multimodal Models (LMMs), to generate parametric pattern-making programs from multi-modal design concepts. LMM offers an intuitive interface for interpreting diverse design inputs, while pattern-making programs could serve as well-structured and semantically meaningful representations of sewing patterns, and act as a robust bridge connecting the cross-domain pattern-making knowledge embedded in LMMs with vectorized sewing patterns. Experimental results demonstrate that our method can flexibly handle various complex design expressions such as images, textual descriptions, designer sketches, or their combinations, and convert them into size-precise sewing patterns with correct stitches. Compared to previous methods, our approach significantly enhances training efficiency, generation quality, and authoring flexibility. Project page: https://style3d.github.io/design2garmentcode.
****

Uncertainty Weighted Gradients for Model Calibration
Jinxu Lin, Linwei Tao, Minjing Dong, Chang Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15497-15507
Model calibration is essential for ensuring that the predictions of deep neural networks accurately reflect true probabilities in real-world classification tasks. However, deep networks often produce over-confident or under-confident predictions, leading to miscalibration. Various methods have been proposed to address this issue by designing effective loss functions for calibration, such as focal loss. In this paper, we analyze its effectiveness and provide a unified loss framework of focal loss and its variants, where we mainly attribute their superiority in model calibration to the loss weighting factor that estimates sample-wise uncertainty. Based on our analysis, existing loss functions fail to achieve optimal calibration performance due to two main issues: including misalignment during optimization and insufficient precision in uncertainty estimation. Specifically, focal loss cannot align sample uncertainty with gradient scaling and the single logit cannot indicate the uncertainty. To address these issues, we reformulate the optimization from the perspective of gradients, which focuses on uncertain

samples. Meanwhile, we propose using the Brier Score as the loss weight factor, which provides a more accurate uncertainty estimation via all the logits. Extensive experiments on various models and datasets demonstrate that our method achieves state-of-the-art (SOTA) performance.

****

Efficient Dynamic Scene Editing via 4D Gaussian-based Static-Dynamic Separation
Joohyun Kwon, Hanbyel Cho, Junmo Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26855-26865

Recent 4D dynamic scene editing methods require editing thousands of 2D images used for dynamic scene synthesis and updating the entire scene with additional training loops, resulting in several hours of processing to edit a single dynamic scene. Therefore, these methods are not scalable with respect to the temporal dimension of the dynamic scene (i.e., the number of timesteps). In this work, we propose Instruct-4DGS, an efficient dynamic scene editing method that is more scalable in terms of temporal dimension. To achieve computational efficiency, we leverage a 4D Gaussian representation that models a 4D dynamic scene by combining static 3D Gaussians with a Hexplane-based deformation field, which captures dynamic information. We then perform editing solely on the static 3D Gaussians, which is the minimal but sufficient component required for visual editing. To resolve the misalignment between the edited 3D Gaussians and the deformation field, which may arise from the editing process, we introduce a refinement stage using a score distillation mechanism. Extensive editing results demonstrate that Instruct-4DGS is efficient, reducing editing time by more than half compared to existing methods while achieving high-quality edits that better follow user instructions.

****

Unlearning through Knowledge Overwriting: Reversible Federated Unlearning via Selective Sparse Adapter
Zhengyi Zhong, Weidong Bao, Ji Wang, Shuai Zhang, Jingxuan Zhou, Lingjuan Lyu, Wei Yang Bryan Lim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30661-30670

Federated Learning is a promising paradigm for privacy-preserving collaborative model training. In practice, it is essential not only to continuously train the model to acquire new knowledge but also to guarantee old knowledge the right to be forgotten (i.e., federated unlearning), especially for privacy-sensitive information or harmful knowledge. However, current federated unlearning methods face several challenges, including indiscriminate unlearning of cross-client knowledge, irreversibility of unlearning, and significant unlearning costs. To this end, we propose a method named FUSED, which first identifies critical layers by analyzing each layer's sensitivity to knowledge and constructs sparse unlearning adapters for sensitive ones. Then, the adapters are trained without altering the original parameters, overwriting the unlearning knowledge with the remaining knowledge. This knowledge overwriting process enables FUSED to mitigate the effects of indiscriminate unlearning. Moreover, the introduction of independent adapters makes unlearning reversible and significantly reduces the unlearning costs. Finally, extensive experiments on five datasets across three unlearning scenarios demonstrate that FUSED's effectiveness is comparable to Retraining, surpassing all other baselines while greatly reducing unlearning costs.

****

SocialMOIF: Multi-Order Intention Fusion for Pedestrian Trajectory Prediction
Kai Chen, Xiaodong Zhao, Yujie Huang, Guoyu Fang, Xiao Song, Ruiping Wang, Ziyuan Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22465-22475

The analysis and prediction of agent trajectories are crucial for decision-making processes in intelligent systems, with precise short-term trajectory forecasting being highly significant across a range of applications. Agents and their social interactions have been quantified and modeled by researchers from various perspectives; however, substantial limitations exist in the current work due to the inherent high uncertainty of agent intentions and the complex higher-order influences among neighboring groups. SocialMOIF is proposed to tackle these challen

ges, concentrating on the higher-order intention interactions among neighboring groups while reinforcing the primary role of first-order intention interactions between neighbors and the target agent. This method develops a multi-order intention fusion model to achieve a more comprehensive understanding of both direct and indirect intention information. Within SocialMOIF, a trajectory distribution approximator is designed to guide the trajectories toward values that align more closely with the actual data, thereby enhancing model interpretability. Furthermore, a global trajectory optimizer is introduced to enable more accurate and efficient parallel predictions. By incorporating a novel loss function that accounts for distance and direction during training, experimental results demonstrate that the model outperforms previous state-of-the-art baselines across multiple metrics in both dynamic and static datasets.
****

FFaceNeRF: Few-shot Face Editing in Neural Radiance Fields
Kwan Yun, Chaelin Kim, Hangyeul Shin, Junyong Noh; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10825-10835
Recent 3D face editing methods using masks have produced high-quality edited images by leveraging Neural Radiance Fields (NeRF). Despite their impressive performance, existing methods often provide limited user control due to the use of pre-trained segmentation masks. To utilize masks with a desired layout, an extensive training dataset is required, which is challenging to gather. We present FFaceNeRF, a NeRF-based face editing technique that can overcome the challenge of limited user control due to the use of fixed mask layouts. Our method employs a geometry adapter with feature injection, allowing for effective manipulation of geometry attributes. Additionally, we adopt latent mixing for tri-plane augmentation, which enables training with a few samples. This facilitates rapid model adaptation to desired mask layouts, crucial for applications in fields like personalized medical imaging or creative face editing. Our comparative evaluations demonstrate that FFaceNeRF surpasses existing mask based face editing methods in terms of flexibility, control, and generated image quality, paving the way for future advancements in customized and high-fidelity 3D face editing.
****

Discrete to Continuous: Generating Smooth Transition Poses from Sign Language Observations
Shengeng Tang, Jiayi He, Lechao Cheng, Jingjing Wu, Dan Guo, Richang Hong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3481-3491
Generating continuous sign language videos from discrete segments is challenging due to the need for smooth transitions that preserve natural flow and meaning. Traditional approaches that simply concatenate isolated signs often result in abrupt transitions, disrupting video coherence. To address this, we propose a novel framework, Sign-D2C, that employs a conditional diffusion model to synthesize contextually smooth transition frames, enabling the seamless construction of continuous sign language sequences. Our approach transforms the unsupervised problem of transition frame generation into a supervised training task by simulating the absence of transition frames through random masking of segments in long-duration sign videos. The model learns to predict these masked frames by denoising Gaussian noise, conditioned on the surrounding sign observations, allowing it to handle complex, unstructured transitions. During inference, we apply a linearly interpolating padding strategy that initializes missing frames through interpolation between boundary frames, providing a stable foundation for iterative refinement by the diffusion model. Extensive experiments on the PHOENIX14T, USTC-CSL100, and USTC-SLR500 datasets demonstrate the effectiveness of our method in producing continuous, natural sign language videos.
****

HistoFS: Non-IID Histopathologic Whole Slide Image Classification via Federated Style Transfer with RoI-Preserving
Farchan Hakim Raswa, Chun-Shien Lu, Jia-Ching Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30251-30260
Federated learning for pathological whole slide image (WSI) classification allow

s multiple clients to train a global multiple instance learning (MIL) model without sharing their privacy-sensitive WSIs. To accommodate the non-independent and identically distributed (non-i.i.d.) feature shifts, cross-client style transfer has been popularly used but is subject to two fundamental issues: (1) WSI contains multiple morphological structures, each corresponding to a distinct style. (2) Performing style transfer may potentially shift the region of interests (RoIs) in the augmented WSIs. To address these challenges, we propose HistoFS, a federated learning framework for computational pathology on non-i.i.d. feature shifts in WSI classification. Specifically, we introduce pseudo bag styles that capture multiple style variations within a single WSI. In addition, an authenticity module is introduced to ensure that RoIs are preserved, allowing local models to learn WSIs with diverse styles while maintaining essential RoIs. Extensive experiments validate the superiority of HistoFS over state-of-the-art methods on three clinical datasets. Our code is available at https://lalakitchen.github.io/HistoFS/.

****

Unified Medical Lesion Segmentation via Self-referring Indicator

Shijie Chang, Xiaoqi Zhao, Lihe Zhang, Tiancheng Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10414-10424

The recently emerged in-context-learning-based (ICL-based) models have the potential towards the unification of medical lesion segmentation. However, due to their cross-fusion designs, existing ICL-based unified segmentation models fail to accurately localize lesions with low-matched reference sets. Considering that the query itself can be regarded as a high-matched reference, which better indicates the target, we design a self-referencing mechanism that adaptively extracts self-referring indicator vectors from the query based on coarse predictions, thus effectively overcoming the negative impact caused by low-match reference sets. To further facilitate the self-referring mechanism, we introduce reference indicator generation to efficiently extract reference information for coarse predictions instead of using cross-fusion modules, which heavily rely on reference sets. Our designs successfully address the challenges of applying ICL to unified medical lesion segmentation, forming a novel framework named SR-ICL. Our method achieves state-of-the-art results on 8 medical lesion segmentation tasks with only 4 image-mask pairs as reference. Notably, SR-ICL still accomplishes remarkable performance even when using weak reference annotations such as boxes and points, and maintains fixed and low memory consumption even if more tasks are combined. We hope that SR-ICL can provide new insights for the clinical application of medical lesion segmentation.

****

SGSST: Scaling Gaussian Splatting Style Transfer

Bruno Galerne, Jianling Wang, Lara Raad, Jean-Michel Morel; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26535-26544

Applying style transfer to a full 3D environment is a challenging task that has seen many developments since the advent of neural rendering. 3D Gaussian splatting (3DGS) has recently pushed further many limits of neural rendering in terms of training speed and reconstruction quality. This work introduces SGSST: Scaling Gaussian Splatting Style Transfer, an optimization-based method to apply style transfer to pretrained 3DGS scenes. We demonstrate that a new multiscale loss based on global neural statistics, that we name SOS for Simultaneously Optimized Scales, enables style transfer to ultra-high resolution 3D scenes. Not only SGSST pioneers 3D scene style transfer at such high image resolutions, it also produces superior visual quality as assessed by thorough qualitative, quantitative and perceptual comparisons.

****

Learning Bijective Surface Parameterization for Inferring Signed Distance Functions from Sparse Point Clouds with Grid Deformation

Takeshi Noda, Chao Chen, Junsheng Zhou, Weiqi Zhang, Yu-Shen Liu, Zhizhong Han; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22139-22149

Inferring signed distance functions (SDFs) from sparse point clouds remains a ch

allenge in surface reconstruction. The key lies in the lack of detailed geometri c information in sparse point clouds, which is essential for learning a continuo us field. To resolve this issue, we present a novel approach that learns a dynam ic deformation network to predict SDFs in an end-to-end manner. To parameterize a continuous surface from sparse points, we propose a bijective surface paramete rization (BSP) that learns the global shape from local patches. Specifically, we construct a bijective mapping for sparse points from the parametric domain to 3 D local patches, integrating patches into the global surface. Meanwhile, we intr oduce grid deformation optimization (GDO) into the surface approximation to opti mize the deformation of grid points and further refine the parametric surfaces. Experimental results on synthetic and real scanned datasets demonstrate that our method significantly outperforms the current state-of-the-art methods.
****

Minimizing Labeled, Maximizing Unlabeled: An Image-Driven Approach for Video Ins tance Segmentation

Fangyun Wei, Jinjing Zhao, Kun Yan, Chang Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19304-19314

Traditional video instance segmentation (VIS) models rely on extensive per-frame video annotations, which are both time-consuming and costly. In this paper, we present MinMaxVIS, a novel VIS framework that reduces the dependency on fully la beled video datasets by utilizing a small set of labeled images from the target domain along with a large volume of general-domain, unlabeled images. MinMaxVIS operates in three stages: first, a preliminary segmentation model is trained on the small labeled set from the target domain; this model then retrieves relevant instances from the unlabeled dataset to build a high-quality pseudo-labeled set , ensuring a rich content alignment with the target domain while avoiding the in efficiencies of large-scale semi-supervised learning across the entire unlabeled set. Finally, we train MinMaxVIS on a combination of labeled and pseudo-labeled data, addressing challenges such as noise in pseudo-labels and instance associa tion across frames. To simulate object continuity, we augment static images to c reate paired frames, allowing MinMaxVIS to capture instance associations effecti vely. MinMaxVIS outperforms the prior image-driven approach, MinVIS, achieving s uperior mAP scores with significantly reduced labeled data. For instance, MinMax VIS with a Swin-L backbone attains 62.2 mAP on YouTube-VIS 2019 using only 2% la beled data and additional unlabeled images from SA-1B. This surpasses MinVIS, wh ich uses the same backbone trained on fully labeled YouTube-VIS 2019, by 0.6 mAP .
****

Layer- and Timestep-Adaptive Differentiable Token Compression Ratios for Efficie nt Diffusion Transformers

Haoran You, Connelly Barnes, Yuqian Zhou, Yan Kang, Zhenbang Du, Wei Zhou, Lingz hi Zhang, Yotam Nitzan, Xiaoyang Liu, Zhe Lin, Eli Shechtman, Sohrab Amirghodsi, Yingyan Celine Lin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18072-18082

Diffusion Transformers (DiTs) have achieved state-of-the-art (SOTA) image genera tion quality but suffer from high latency and memory inefficiency, making them d ifficult to deploy on resource-constrained devices. One major efficiency bottlen eck is that existing DiTs apply equal computation across all regions of an image . However, not all image tokens are equally important, and certain localized are as require more computation, such as objects. To address this, we propose DiffCR , a dynamic DiT inference framework with differentiable compression ratios, whic h automatically learns to dynamically route computation across layers and timest eps for each image token, resulting in efficient DiTs. Specifically, DiffCR inte grates three features: (1) A token-level routing scheme where each DiT layer inc ludes a router that is fine-tuned jointly with model weights to predict token im portance scores. In this way, unimportant tokens bypass the entire layer's compu tation; (2) A layer-wise differentiable ratio mechanism where different DiT laye rs automatically learn varying compression ratios from a zero initialization, re sulting in large compression ratios in redundant layers while others remain less compressed or even uncompressed; (3) A timestep-wise differentiable ratio mecha

nism where each denoising timestep learns its own compression ratio. The resulting pattern shows higher ratios for noisier timesteps and lower ratios as the image becomes clearer. Extensive experiments on text-to-image and inpainting tasks show that DiffCR effectively captures dynamism across token, layer, and timestep axes, achieving superior trade-offs between generation quality and efficiency compared to prior works. The project website is available at https://www.haoranyou.com/diffcr.

****

Zero-shot RGB-D Point Cloud Registration with Pre-trained Large Vision Model
Haobo Jiang, Jin Xie, Jian Yang, Liang Yu, Jianmin Zheng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16943-16952
This paper introduces ZeroMatch, a novel zero-shot RGB-D point cloud registration framework, aimed at achieving robust 3D matching on unseen data without any task-specific training. Our core idea is to utilize the powerful zero-shot image representation of Stable Diffusion, achieved through extensive pre-training on large-scale data, to enhance point-cloud geometric descriptors for robust matching. Specifically, we combine the handcrafted geometric descriptor FPFH with Stable-Diffusion features to create point descriptors that are both locally and contextually aware, enabling reliable RGB-D registration with zero-shot capability. This approach is based on our observation that Stable-Diffusion features effectively encode discriminative global contextual cues, naturally alleviating the feature ambiguity that FPFH often encounters in scenes with repetitive patterns or low overlap. To further enhance cross-view consistency of Stable-Diffusion features for improved matching, we propose a coupled-image input mode that concatenates the source and target images into a single input, replacing the original single-image mode. This design achieves both inter-image and prompt-to-image consistency attentions, facilitating robust cross-view feature interaction and alignment. Finally, we leverage feature nearest neighbors to construct putative correspondences for hypothesize-and-verify transformation estimation. Extensive experiments on 3DMatch, ScanNet, and ScanLoNet verify the excellent zero-shot matching ability of our method.

****

Balancing Two Classifiers via A Simplex ETF Structure for Model Calibration
Jiani Ni, He Zhao, Jintong Gao, Dandan Guo, Hongyuan Zha; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30712-30721
In recent years, deep neural networks (DNNs) have demonstrated state-of-the-art performance across various domains. However, despite their success, they often face calibration issues, particularly in safety-critical applications such as autonomous driving and healthcare, where unreliable predictions can have serious consequences. Recent research starts to improve model calibration from the view of classifier. However, the explore about designing the classifier to solve the model calibration problem is insufficient. Let alone most of existing methods ignore the calibration errors arising from underconfidence. In this work, we propose a novel method by Balancing learnable and ETF classifiers to solve the overconfidence or underconfidence problem for model Calibration named BalCAL. By introducing a confidence-tunable module and a dynamic adjustment method, we ensure better alignment between model confidence and its true accuracy. Extensive experimental validation shows that ours significantly improves model calibration performance while maintaining high predictive accuracy, outperforming existing techniques. This provides a novel solution to the calibration challenges commonly encountered in deep learning. Our code is available at https://github.com/JianiNi/BalCAL.

****

DistinctAD: Distinctive Audio Description Generation in Contexts
Bo Fang, Wenhao Wu, Qiangqiang Wu, Yuxin Song, Antoni B. Chan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13571-13581
Audio Descriptions (ADs) aim to provide a narration of a movie in text form, describing non-dialogue-related narratives, such as characters, actions, or scene establishment. Automatic generation of ADs remains challenging due to: i) the dom

ain gap between movie-AD data and existing data used to train vision-language mo
dels, and ii) the issue of contextual redundancy arising from highly similar nei
ghboring visual clips in a long movie. In this work, we propose **DistinctAD**,
a novel two-stage framework for generating ADs that emphasize distinctiveness to
 produce better narratives. To address the domain gap, we introduce a CLIP-AD ad
aptation strategy that does not require additional AD corpora, enabling more eff
ective alignment between movie and AD modalities at both global and fine-grained
 levels. In Stage-II, DistinctAD incorporates two key innovations: (i) a Context
ual Expectation-Maximization Attention (EMA) module that reduces redundancy by e
xtracting common bases from consecutive video clips, and (ii) an explicit distin
ctive word prediction loss that filters out repeated words in the context, ensur
ing the prediction of unique terms specific to the current AD. Comprehensive eva
luations on MAD-Eval, CMD-AD, and TV-AD benchmarks demonstrate the superiority o
f DistinctAD, with the model consistently outperforming baselines, particularly
in Recall@k/N, highlighting its effectiveness in producing high-quality, distinc
tive ADs.
****

DAMM-Diffusion: Learning Divergence-Aware Multi-Modal Diffusion Model for Nanopa
rticles Distribution Prediction

Junjie Zhou, Shouju Wang, Yuxia Tang, Qi Zhu, Daoqiang Zhang, Wei Shao; Proceedi
ngs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp.
30886-30895

The prediction of nanoparticles (NPs) distribution is crucial for the diagnosis
and treatment of tumors. Recent studies indicate that the heterogeneity of tumor
 microenvironment (TME) highly affects the distribution of NPs across tumors. He
nce, it has become a research hotspot to generate the NPs distribution by the ai
d of multi-modal TME components. However, the distribution divergence among mult
i-modal TME components may cause side effects i.e., the best uni-modal model may
 outperform the joint generative model. To address the above issues, we propose
a \Divergence-Aware Multi-Modal Diffusion model (i.e., DAMM-Diffusion) to adapti
vely generate the prediction results from uni-modal and multi-modal branches in
a unified network. In detail, the uni-modal branch is composed of the U-Net arch
itecture while the multi-modal branch extends it by introducing two novel fusion
 modules i.e., Multi-Modal Fusion Module (MMFM) and Uncertainty-Aware Fusion Mod
ule (UAFM). Specifically, the MMFM is proposed to fuse features from multiple mo
dalities, while the UAFM module is introduced to learn the uncertainty map for c
ross-attention computation. Following the individual prediction results from eac
h branch, the Divergence-Aware Multi-Modal Predictor (DAMMP) module is proposed
to assess the consistency of multi-modal data with the uncertainty map, which de
termines whether the final prediction results come from multi-modal or uni-moda
l predictions. We predict the NPs distribution given the TME components of tumor
 vessels and cell nuclei, and the experimental results show that DAMM-Diffusion
can generate the distribution of NPs with higher accuracy than the comparing met
hods. Additional results on the multi-modal brain image synthesis task further v
alidate the effectiveness of the proposed method.
****

Unveiling Differences in Generative Models: A Scalable Differential Clustering A
pproach

Jingwei Zhang, Mohammad Jalali, Cheuk Ting Li, Farzan Farnia; Proceedings of the
 Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8269-8278

A fine-grained comparison of generative models requires the identification of sa
mple types generated differently by each of the involved models. While quantitat
ive scores have been proposed in the literature to rank different generative mod
els, score-based evaluation and ranking do not reveal the nuanced differences be
tween the generative models in producing different sample types. In this work, w
e propose solving a differential clustering problem to detect sample types gener
ated differently by two generative models. To solve the differential clustering
problem, we develop a spectral method called Fourier-based Identification of Nov
el Clusters (FINC) to identify modes produced by a generative model with a highe
r frequency in comparison to a reference distribution. FINC provides a scalable

algorithm based on random Fourier features to estimate the eigenspace of kernel covariance matrices of two generative models and utilize the principal eigendirections to detect the sample types present more dominantly in each model. We demonstrate the application of the FINC method to large-scale computer vision datasets and generative modeling frameworks. Our numerical results suggest the scalability of the developed Fourier-based method in highlighting the sample types produced with different frequencies by generative models. The project code is available at https://github.com/buyeah1109/FINC

****

CL-MoE: Enhancing Multimodal Large Language Model with Dual Momentum Mixture-of-Experts for Continual Visual Question Answering

Tianyu Huai, Jie Zhou, Xingjiao Wu, Qin Chen, Qingchun Bai, Ze Zhou, Liang He; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19608-19617

Multimodal large language models (MLLMs) have garnered widespread attention from researchers due to their remarkable understanding and generation capabilities in visual language tasks (e.g., visual question answering). However, the rapid pace of knowledge updates in the real world makes offline training of MLLMs costly, and when faced with non-stationary data streams, MLLMs suffer from catastrophic forgetting during learning. In this paper, we propose an MLLMs-based dual momentum Mixture-of-Experts (\texttt CL-MoE) framework for continual visual question answering. We integrate MLLMs with continual learning to utilize the rich commonsense knowledge in LLMs.We introduce a Dual-Router MoE (RMoE) to select the global and local experts using task-level and instance-level routers, to robustly assign weights to the experts most appropriate for the task. Then, we design a dynamic Momentum MoE (MMoE) to update the parameters of experts dynamically based on the relationships between the experts and tasks, so that the model can absorb new knowledge while maintaining existing knowledge. The extensive experimental results indicate that our method achieves state-of-the-art performance on 10 VQA tasks, proving the effectiveness of our approach. The codes and weights will be released on GitHub.

****

Semantic Library Adaptation: LoRA Retrieval and Fusion for Open-Vocabulary Semantic Segmentation

Reza Qorbani, Gianluca Villani, Theodoros Panagiotakopoulos, Marc Botet Colomer, Linus Härenstam-Nielsen, Mattia Segu, Pier Luigi Dovesi, Jussi Karlgren, Daniel Cremers, Federico Tombari, Matteo Poggi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9804-9815

Open-vocabulary semantic segmentation models associate vision and text to label pixels from an undefined set of classes using textual queries, providing versatile performance on novel datasets. However, large shifts between training and test domains degrade their performance, requiring fine-tuning for effective real-world applications. We introduce Semantic Library Adaptation (SemLA), a novel framework for training-free, test-time domain adaptation. SemLA leverages a library of LoRA-based adapters indexed with CLIP embeddings, dynamically merging the most relevant adapters based on proximity to the target domain in the embedding space. This approach constructs an ad-hoc model tailored to each specific input without additional training. Our method scales efficiently, enhances explainability by tracking adapter contributions, and inherently protects data privacy, making it ideal for sensitive applications. Comprehensive experiments on a 20-domain benchmark built over 10 standard datasets demonstrate SemLA's superior adaptability and performance across diverse settings, establishing a new standard in domain adaptation for open-vocabulary semantic segmentation.

****

PhyS-EdiT: Physics-aware Semantic Image Editing with Text Description

Ziqi Cai, Shuchen Weng, Yifei Xia, Boxin Shi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7867-7876

Achieving joint control over material properties, lighting, and high-level semantics in images is essential for applications in digital media, advertising, and interactive design. Existing methods often isolate these properties, lacking a c

ohesive approach to manipulating materials, lighting, and semantics simultaneous ly. We introduce PhyS-EdiT, a novel diffusion-based model that enables precise c ontrol over four critical material properties: roughness, metallicity, albedo, a nd transparency while integrating lighting and semantic adjustments within a sin gle framework. To facilitate this disentangled control, we present PR-TIPS, a la rge and diverse synthetic dataset designed to improve the disentanglement of mat erial and lighting effects. PhyS-EdiT incorporates a dual-network architecture a nd robust training strategies to balance low-level physical realism with high-le vel semantic coherence, supporting localized and continuous property adjustments . Extensive experiments demonstrate the superiority of PhyS-EdiT in editing both synthetic and real-world images, achieving state-of-the-art performance on mate rial, lighting, and semantic editing tasks.

****

U-Know-DiffPAN: An Uncertainty-aware Knowledge Distillation Diffusion Framework with Details Enhancement for PAN-Sharpening

Sungpyo Kim, Jeonghyeok Do, Jaehyup Lee, Munchurl Kim; Proceedings of the Comput er Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23069-23079

Conventional methods for PAN-sharpening often struggle to restore fine details d ue to limitations in leveraging high-frequency information. Moreover, diffusion- based approaches lack sufficient conditioning to fully utilize Panchromatic (PAN) images and low-resolution multispectral (LRMS) inputs effectively. To address these challenges, we propose an uncertainty-aware knowledge distillation diffusi on framework with details enhancement for PAN-sharpening, called U-Know-DiffPAN. The U-Know-DiffPAN incorporates uncertainty-aware knowledge distillation for ef fective transfer of feature details from our teacher model to a student one. The teacher model in our U-Know-DiffPAN captures frequency details through freqeunc y selective attention, facilitating accurate reverse process learning. By condit ioning the encoder on compact vector representations of PAN and LRMS and the dec oder on Wavelet transforms, we enable rich frequency utilization. So, the high-c apacity teacher model distills frequency-rich features into a lightweight studen t model aided by an uncertainty map. From this, the teacher model can guide the student model to focus on difficult image regions for PAN-sharpening via the usa ge of the uncertainty map. Extensive experiments on diverse datasets demonstrate the robustness and superior performance of our U-Know-DiffPAN over very recent state-of-the-art PAN-sharpening methods. The project page is available at https: //kaist-viclab.github.io/U-Know-DiffPAN-site/ https://kaist-viclab.github.io/U- Know-DiffPAN-site

****

SceneDiffuser++: City-Scale Traffic Simulation via a Generative World Model

Shuhan Tan, John Lambert, Hong Jeon, Sakshum Kulshrestha, Yijing Bai, Jing Luo, Dragomir Anguelov, Mingxing Tan, Chiyu Max Jiang; Proceedings of the Computer Vi sion and Pattern Recognition Conference (CVPR), 2025, pp. 1570-1580

The goal of traffic simulation is to augment a potentially limited amount of man ually-driven miles that is available for testing and validation, with a much lar ger amount of simulated synthetic miles. The culmination of this vision would be a generative simulated city, where given a map of the city and an autonomous ve hicle (AV) software stack, the simulator can seamlessly simulate the trip from p oint A to point B by populating the city around the AV and controlling all aspec ts of the scene, from animating the dynamic agents (e.g., vehicles, pedestrians) to controlling the traffic light states. We refer to this vision as CitySim, wh ich requires an agglomeration of simulation technologies: scene generation to po pulate the initial scene, agent behavior modeling to animate the scene, occlusio n reasoning, dynamic scene generation to seamlessly spawn and remove agents, and environment simulation for factors such as traffic lights. While some key techn ologies have been separately studied in various works, others such as dynamic sc ene generation and environment simulation have received less attention in the re search community. We propose SceneDiffuser++, the first end-to-end generative wo rld model trained on a single loss function capable of point A-to-B simulation o n a city scale integrating all the requirements above. We demonstrate the city-s cale traffic simulation capability of SceneDiffuser++ and study its superior rea

lism under long simulation conditions.We evaluate the simulation quality on an augmented version of the Waymo Open Motion Dataset (WOMD) with larger map regions to support trip-level simulation.
****

Gaussian Splatting Feature Fields for (Privacy-Preserving) Visual Localization
Maxime Pietrantoni, Gabriela Csurka, Torsten Sattler; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1082-1092
Visual localization is the task of estimating a camera pose in a known environment. In this paper, we utilize 3D Gaussian Splatting (3DGS)-based representations for accurate and privacy-preserving visual localization. We propose Gaussian Splatting Feature Fields (GSFFs), a scene representation for visual localization that combines an explicit geometry model (3DGS) with an implicit feature field. We leverage the dense geometric information and differentiable rasterization algorithm from 3DGS to learn robust feature representations grounded in 3D. In particular, we align a 3D scale-aware feature field and a 2D feature encoder in a common embedding space through a contrastive framework. Using a 3D structure-informed clustering procedure, we further regularize the representation learning and seamlessly convert the features to segmentations, which can be used for privacy-preserving visual localization. Pose refinement, which involves aligning either feature maps or segmentations from a query image with those rendered from the GSFFs scene representation, is used to achieve localization. The resulting privacy- and non-privacy-preserving localization pipelines, evaluated on multiple real-world datasets, show state-of-the-art performances.
****

Point Cloud Upsampling Using Conditional Diffusion Module with Adaptive Noise Suppression
Boqian Zhang, Shen Yang, Hao Chen, Chao Yang, Jing Jia, Guang Jiang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16987-16996
Point cloud upsampling can improve the quality of the initial point cloud, significantly enhancing the performance of downstream tasks such as classification and segmentation. Existing methods mostly focus on generating the geometric details of point clouds, neglecting noise suppression. To address this, we propose a novel network based on a conditional diffusion model, incorporating the Adaptive Noise Suppression (ANS) module, which we refer to as PDANS. The ANS module assigns weights to each point and determines the removal strategy based on these weights, reducing the impact of noisy points on the sampling process. The module first selects the neighborhood set for each point in the point cloud and performs a weighted sum between the point and its neighbors. It then adjusts the removal points based on the weighted sum, effectively mitigating the bias caused by outliers. We introduce the TreeTrans (TT) module to capture more correlated feature information. This module learns the interaction between high-level and low-level features, resulting in a more comprehensive and refined feature representation. Our results on several widely used benchmark datasets demonstrate that PDANS exhibits exceptional robustness in noisy point cloud processing and outperforms current state-of-the-art(SOTA) methods in terms of performance. Code is available at https://github.com/Baty2023/PDANS.
****

Gazing Into Missteps: Leveraging Eye-Gaze for Unsupervised Mistake Detection in Egocentric Videos of Skilled Human Activities
Michele Mazzamuto, Antonino Furnari, Yoichi Sato, Giovanni Maria Farinella; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8310-8320
We address the challenge of unsupervised mistake detection in egocentric video of skilled human activities through the analysis of gaze signals. While traditional methods rely on manually labeled mistakes, our approach does not require mistake annotations, hence overcoming the need of domain-specific labeled data. Based on the observation that eye movements closely follow object manipulation activities, we assess to what extent eye-gaze signals can support mistake detection, proposing to identify deviations in attention patterns measured through a gaze t

racker with respect to those estimated by a gaze prediction model. Since predict ing gaze in video is characterized by high uncertainty, we propose a novel gaze completion task, where eye fixations are predicted from visual observations and partial gaze trajectories, and contribute a novel gaze completion approach which explicitly models correlations between gaze information and local visual tokens . Inconsistencies between predicted and observed gaze trajectories act as an ind icator to identify mistakes. Experiments highlight the effectiveness of the prop osed approach in different settings, with relative gains up to +14%, +11%, and +5% in EPIC-Tent, HoloAssist and IndustReal respectively, remarkably matching re sults of supervised approaches without seeing any labels. We further show that g aze-based analysis is particularly useful in the presence of skilled actions, lo w action execution confidence, and actions requiring hand-eye coordination and o bject manipulation skills. Our method is ranked first on the HoloAssist Mistake Detection challenge.

****

Trajectory Mamba: Efficient Attention-Mamba Forecasting Model Based on Selective SSM

Yizhou Huang, Yihua Cheng, Kezhi Wang; Proceedings of the Computer Vision and Pa ttern Recognition Conference (CVPR), 2025, pp. 12058-12067

Motion prediction is crucial for autonomous driving, as it enables accurate fore casting of future vehicle trajectories based on historical inputs. This paper in troduces Trajectory Mamba, a novel efficient trajectory prediction framework bas ed on the selective state-space model (SSM). Conventional attention-based models face the challenge of computational costs that grow quadratically with the numb er of targets, hindering their application in highly dynamic environments. In re sponse, we leverage the SSM to redesign the self-attention mechanism in the enco der-decoder architecture, thereby achieving linear time complexity. To address t he potential reduction in prediction accuracy resulting from modifications to th e attention mechanism, we propose a joint polyline encoding strategy to better c apture the associations between static and dynamic contexts, ultimately enhancin g prediction accuracy. Additionally, to balance prediction accuracy and inferenc e speed, we adopted the decoder that differs entirely from the encoder. Through cross-state space attention, all target agents share the scene context, allowing the SSM to interact with the shared scene representation during decoding, thus inferring different trajectories over the next prediction steps. Our model achie ves state-of-the-art results in terms of inference speed and parameter efficienc y on both the Argoverse 1 and Argoverse 2 datasets. It demonstrates a four-fold reduction in FLOPs compared to existing methods and reduces parameter count by o ver 40% while surpassing the performance of the vast majority of previous method s. These findings validate the effectiveness of Trajectory Mamba in trajectory p rediction tasks.

****

Recover and Match: Open-Vocabulary Multi-Label Recognition through Knowledge-Con strained Optimal Transport

Hao Tan, Zichang Tan, Jun Li, Ajian Liu, Jun Wan, Zhen Lei; Proceedings of the C omputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4650-4660

Identifying multiple novel classes in an image, known as open-vocabulary multi-l abel recognition, is a challenging task in computer vision. Recent studies explo re the transfer of powerful vision-language models such as CLIP. However, these approaches face two critical challenges: (1) The local semantics of CLIP are dis rupted due to its global pre-training objectives, resulting in unreliable region al predictions. (2) The matching property between image regions and candidate la bels has been neglected, relying instead on naive feature aggregation such as av erage pooling, which leads to spurious predictions from irrelevant regions. In t his paper, we present RAM (Recover And Match), a novel framework that effectivel y addresses the above issues. To tackle the first problem, we propose Ladder Loc al Adapter (LLA) to enforce the model to refocus on its surrounding local region s, recovering local semantics in a memory-friendly way. For the second issue, we propose Knowledge-Constrained Optimal Transport (KCOT) to suppress meaningless matching to non-GT labels by formulating the task as an optimal transport proble

m. As a result, RAM achieves state-of-the-art performance on various datasets fr
om three distinct domains, and shows great potential to boost the existing metho
ds. Code: https://github.com/EricTan7/RAM.
****

RelationField: Relate Anything in Radiance Fields
Sebastian Koch, Johanna Wald, Mirco Colosi, Narunas Vaskevicius, Pedro Hermosill
a, Federico Tombari, Timo Ropinski; Proceedings of the Computer Vision and Patte
rn Recognition Conference (CVPR), 2025, pp. 21706-21716
Neural radiance fields are an emerging 3D scene representation and recently even
 been extended to learn features for scene understanding by distilling open-voca
bulary features from vision-language models. However, current method primarily f
ocus on object-centric representations, supporting object segmentation or detect
ion, while understanding semantic relationships between objects remains largely
unexplored. To address this gap, we propose RelationField, the first method to e
xtract inter-object relationships directly from neural radiance fields. Relation
Field represents relationships between objects as pairs of rays within a neural
radiance field, effectively extending its formulation to include implicit relati
onship queries. To teach RelationField complex, open-vocabulary relationships, r
elationship knowledge is distilled from multi-modal LLMs. To evaluate RelationFi
eld, we solve open-vocabulary 3D scene graph generation tasks and relationship-g
uided instance segmentation, achieving state-of-the-art performance in both task
s. See the project website at relationfield.github.io.
****

DyFo: A Training-Free Dynamic Focus Visual Search for Enhancing LMMs in Fine-Gra
ined Visual Understanding
Geng Li, Jinglin Xu, Yunzhen Zhao, Yuxin Peng; Proceedings of the Computer Visio
n and Pattern Recognition Conference (CVPR), 2025, pp. 9098-9108
Humans can effortlessly locate desired objects in cluttered environments, relyin
g on a cognitive mechanism known as visual search to efficiently filter out irre
levant information and focus on task related regions. Inspired by this process,
we propose DyFo (Dynamic Focus), a training-free dynamic focusing visual search
method that enhances fine-grained visual understanding in large multi-modal mode
ls (LMMs). Unlike existing approaches which require additional modules or data c
ollection, DyFo leverages a bidirectional interaction between LMMs and visual ex
perts, using a Monte Carlo Tree Search (MCTS) algorithm to simulate human-like f
ocus adjustments. This enables LMMs to focus on key visual regions while filteri
ng out irrelevant content, without introducing additional training caused by voc
abulary expansion or the integration of specialized localization modules. Experi
mental results demonstrate that DyFo significantly improves fine-grained visual
understanding and reduces hallucination issues in LMMs, achieving superior perfo
rmance across both fixed and dynamic resolution models.
****

From Head to Tail: Towards Balanced Representation in Large Vision-Language Mode
ls through Adaptive Data Calibration
Mingyang Song, Xiaoye Qu, Jiawei Zhou, Yu Cheng; Proceedings of the Computer Vis
ion and Pattern Recognition Conference (CVPR), 2025, pp. 9434-9444
Large Vision-Language Models (LVLMs) have achieved significant progress in combi
ning visual comprehension with language generation.Despite this success, the tra
ining data of LVLMs still suffers from Long-Tail (LT) problems, where the data d
istribution is highly imbalanced.Previous works have mainly focused on tradition
al VLM architectures, i.e., CLIP or ViT, and specific tasks such as recognition
and classification. Nevertheless, the exploration of LVLM (e.g. LLaVA) and more
general tasks (e.g. Visual Question Answering and Visual Reasoning) remains unde
r-explored.In this paper, we first conduct an in-depth analysis of the LT issues
 in LVLMs and identify two core causes: the overrepresentation of head concepts
and the underrepresentation of tail concepts.Based on the above observation, we
propose an Adaptive Data Refinement Framework (ADR), which consists of two stage
s: Data Rebalancing (DR) and Data Synthesis (DS).In the DR stage, we adaptively
rebalance the redundant data based on entity distributions, while in the DS stag
e, we leverage Denoising Diffusion Probabilistic Models (DDPMs) and scarce image

s to supplement underrepresented portions.Through comprehensive evaluations acro ss eleven benchmarks, our proposed ADR effectively mitigates the long-tail probl em in the training data, improving the average performance of LLaVA 1.5 relative ly by 4.36%, without increasing the training data volume. Our code and data will be publicly released.
****

Let Humanoids Hike! Integrative Skill Development on Complex Trails
Kwan-Yee Lin, Stella X. Yu; Proceedings of the Computer Vision and Pattern Recog nition Conference (CVPR), 2025, pp. 22498-22507
Hiking on complex trails demands balance, agility, and adaptive decision-making over unpredictable terrain. Current humanoid research remains fragmented and ina dequate for hiking: locomotion focuses on motor skills without long-term goals o r situational awareness, while semantic navigation overlooks real-world embodime nt and local terrain variability. We propose training humanoids to hike on comp lex trails, fostering integrative skill development across visual perception, de cision making, and motor execution. We develop LEGO-H, a learning framework that enables a humanoid with vision to hike complex trails independently. It has two key innovations. (1) A Temporal Vision Transformer anticipates future steps to guide locomotion, unifying local movement and goal-directed navigation. (2) Late nt representations of joint movement patterns combined with hierarchical metric learning allow smooth policy transfer from privileged training to real-world tra ining. These techniques enable LEGO-H to handle diverse physical and environmen tal challenges without relying on predefined motion patterns. Experiments on div erse simulated hiking trails and humanoids with different morphologies demonstra te LEGO-H's robustness and versatility, establishing a strong foundation for fut ure humanoid development.
****

VLOGGER: Multimodal Diffusion for Embodied Avatar Synthesis
Enric Corona, Andrei Zanfir, Eduard Gabriel Bazavan, Nikos Kolotouros, Thiemo Al ldieck, Cristian Sminchisescu; Proceedings of the Computer Vision and Pattern Re cognition Conference (CVPR), 2025, pp. 15896-15908
We propose VLOGGER, a method for audio-driven human video generation from a sing le input image of a person, which builds on the success of recent generative dif fusion models. Our method consists of 1) a stochastic human-to-3d-motion diffusi on model, and 2) a novel diffusion-based architecture that augments text-to-imag e models with both spatial and temporal controls. This supports the generation o f high quality video of variable length, easily controllable through text or spe ech via high-level representations of human faces and bodies. In contrast to pre vious work, our method does not require training for each person, does not rely on face detection and cropping, generates the complete image (not just the face or the lips), and considers a broad spectrum of scenarios (e.g. visible torso or diverse subject identities) that are critical to correctly synthesize humans wh o communicate. We also curate MENTOR, a new and diverse dataset with 3d pose and expression anno- tations, one order of magnitude larger than previous ones (800 ,000 identities) and with dynamic gestures, where we train and ablate our main t echnical contributions. VLOGGER outperforms state-of-the-art methods in three pu blic benchmarks, considering image quality, identity preservation and temporal c onsistency while also generating upperbody gestures. We analyze the performance of VLOGGER with respect to multiple diversity metrics, showing that our architec tural choices and the use of MENTOR benefit training a fair and unbiased model a t scale. Finally we show applications in video editing and personalization.
****

DEIM: DETR with Improved Matching for Fast Convergence
Shihua Huang, Zhichao Lu, Xiaodong Cun, Yongjun Yu, Xiao Zhou, Xi Shen; Proceedi ngs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15162-15171
We introduce DEIM, an innovative and efficient training framework designed to ac celerate convergence in real-time object detection with Transformer-based archit ectures (DETR). To mitigate the sparse supervision inherent in one-to-one (O2O) matching in DETR models, DEIM employs a Dense O2O matching strategy. This approa

ch increases the number of positive samples per image by incorporating additiona l targets, using standard data augmentation techniques. While Dense O2O matching speeds up convergence, it also introduces numerous low-quality matches that cou ld affect performance. To address this, we propose the Matchability-Aware Loss (MAL), a novel loss function that optimizes matches across various quality levels , enhancing the effectiveness of Dense O2O. Extensive experiments on the COCO dataset validate the efficacy of DEIM. When integrated with RT-DETR and D-FINE, it consistently boosts performance while reducing training time by 50%. Notably , paired with RT-DETRv2, DEIM achieves 53.2% AP in a single day of training on a n NVIDIA 4090 GPU. Additionally, DEIM-trained real-time models outperform leadin g real-time object detectors, with DEIM-D-FINE-L and DEIM-D-FINE-X achieving 54. 7% and 56.4% AP at 124 and 78 FPS on an NVIDIA T4 GPU, respectively, without the need for additional data. We believe DEIM sets a new baseline for advancements in real-time object detection. Our code and pre-trained models are available at https://www.shihuahuang.cn/DEIM/.

****

BF-STVSR: B-Splines and Fourier---Best Friends for High Fidelity Spatial-Tempora l Video Super-Resolution

Eunjin Kim, Hyeonjin Kim, Kyong Hwan Jin, Jaejun Yoo; Proceedings of the Compute r Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28009-28018

While prior methods in Continuous Spatial-Temporal Video Super-Resolution (C-STV SR) employ Implicit Neural Representation (INR) for continuous encoding, they of ten struggle to capture the complexity of video data, relying on simple coordina te concatenation and pre-trained optical flow networks for motion representation . Interestingly, we find that adding position encoding, contrary to common obse rvations, does not improve---and even degrades---performance. This issue becomes particularly pronounced when combined with pre-trained optical flow networks, w hich can limit the model's flexibility. To address these issues, we propose BF-S TVSR, a C-STVSR framework with two key modules tailored to better represent spat ial and temporal characteristics of video: 1) B-spline Mapper for smooth tempora l interpolation, and 2) Fourier Mapper for capturing dominant spatial frequencie s. Our approach achieves state-of-the-art in various metrics, including PSNR and SSIM, showing enhanced spatial details and natural temporal consistency. Our co de is available https://github.com/Eunjnnn/bfstvsr.

****

DIO: Decomposable Implicit 4D Occupancy-Flow World Model

Christopher Diehl, Quinlan Sykora, Ben Agro, Thomas Gilles, Sergio Casas, Raquel Urtasun; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27456-27466

We present DIO, a flexible world model that can estimate the scene occupancy-flo w from a sparse set of LiDAR observations, and decompose it into individual inst ances. DIO can not only complete instance shapes at the present time, but also f orecast their occupancy-flow evolution over a future horizon. Thanks to its flex ible prompt representation, DIO can take instance prompts from off-the-shelf mod els like 3D detectors, achieving state-of-the-art performance in the task of 4D semantic occupancy completion and forecasting on the Argoverse 2 dataset. Moreov er, our world model can easily and effectively be transferred to downstream task s like LiDAR point cloud forecasting, ranking first compared to all baselines in the Argoverse 4D occupancy forecasting challenge.

****

SLADE: Shielding against Dual Exploits in Large Vision-Language Models

Md Zarif Hossain, Ahmed Imteaj; Proceedings of the Computer Vision and Pattern R ecognition Conference (CVPR), 2025, pp. 24244-24254

Large Vision-Language Models (LVLMs) have emerged as transformative tools in mul timodal tasks, seamlessly integrating pretrained vision encoders to align visual and textual modalities. Prior works have highlighted the susceptibility of LVLM s to dual exploits (gradient-based and optimization-based jailbreak attacks), wh ich leverage the expanded attack surface introduced by the image modality. Despi te advancements in enhancing robustness, existing methods fall short in their ab ility to defend against dual exploits while preserving fine-grained semantic det

ails and overall semantic coherence under intense adversarial perturbations. To bridge this gap, we introduce SheiLding against Dual Exploits (SLADE), a novel unsupervised adversarial fine-tuning scheme that enhances the resilience of CLIP-based vision encoders. SLADE's dual-level contrastive learning approach balances the granular and the holistic, capturing fine-grained image details without losing sight of high-level semantic coherence. Extensive experiments demonstrate that SLADE-equipped LVLMs set a new benchmark for robustness against dual exploits while preserving fine-grained semantic details of perturbed images. Notably, SLADE achieves these results without compromising the core functionalities of LVLMs, such as instruction following, or requiring the computational overhead (e.g., large batch sizes, momentum encoders) commonly associated with traditional contrastive learning methods.

****

Human Motion Instruction Tuning

Lei Li, Sen Jia, Jianhao Wang, Zhongyu Jiang, Feng Zhou, Ju Dai, Tianfang Zhang, Zongkai Wu, Jenq-Neng Hwang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17582-17591

This paper presents LLaMo (Large Language and Human Motion Assistant), a multimodal framework for human motion instruction tuning. In contrast to conventional instruction-tuning approaches that convert non-linguistic inputs, such as video or motion sequences, into language tokens, LLaMo retains motion in its native form for instruction tuning. This method preserves motion-specific details that are often diminished in tokenization, thereby improving the model's ability to interpret complex human behaviors. By processing both video and motion data alongside textual inputs, LLaMo enables a flexible, human-centric analysis. Experimental evaluations across high-complexity domains, including human behaviors and professional activities, indicate that LLaMo effectively captures domain-specific knowledge, enhancing comprehension and prediction in motion-intensive scenarios. We hope LLaMo offers a foundation for future multimodal AI systems with broad applications, from sports analytics to behavioral prediction.

****

A Flag Decomposition for Hierarchical Datasets

Nathan Mankovich, Ignacio Santamaria, Gustau Camps-Valls, Tolga Birdal; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18738-18748

Flag manifolds encode nested sequences of subspaces and serve as powerful structures for various computer vision and machine learning applications. Despite their utility in tasks such as dimensionality reduction, motion averaging, and subspace clustering, current applications are often restricted to extracting flags using common matrix decomposition methods like the singular value decomposition. Here, we address the need for a general algorithm to factorize and work with hierarchical datasets. In particular, we propose a novel, flag-based method that decomposes arbitrary hierarchical real-valued data into a hierarchy-preserving flag representation in Stiefel coordinates. Our work harnesses the potential of flag manifolds in applications including denoising, clustering, and few-shot learning.

****

RCP-Bench: Benchmarking Robustness for Collaborative Perception Under Diverse Corruptions

Shihang Du, Sanqing Qu, Tianhang Wang, Xudong Zhang, Yunwei Zhu, Jian Mao, Fan Lu, Qiao Lin, Guang Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11908-11918

Collaborative perception enhances single-vehicle perception by integrating sensory data from multiple connected vehicles. However, existing studies often assume ideal conditions, overlooking resilience to real-world challenges such as adverse weather and sensor malfunctions, which is critical for safe deployment. To address this gap, we introduce RCP-Bench, the first comprehensive benchmark designed to evaluate the robustness of collaborative detection models under a wide range of real-world corruptions. RCP-Bench includes three new datasets (i.e., OPV2V-C, V2XSet-C, and DAIR-V2X-C) that simulate six collaborative cases and 14 types

of camera corruption resulting from external environmental factors, sensor failures, and temporal misalignments. Extensive experiments on 10 leading collaborative perception models reveal that, while these models perform well under ideal conditions, they are significantly affected by corruptions. To improve robustness, we propose two simple yet effective strategies, RCP-Drop and RCP-Mix, based on training regularization and feature augmentation. Additionally, we identify several critical factors influencing robustness, such as backbone architecture, camera number, feature fusion methods, and the number of connected vehicles. We hope that RCP-Bench, along with these strategies and insights, will stimulate future research toward developing more robust collaborative perception models. Our benchmark toolkit is available at https://github.com/LuckyDush/RCP-Bench.
****

Olympus: A Universal Task Router for Computer Vision Tasks
Yuanze Lin, Yunsheng Li, Dongdong Chen, Weijian Xu, Ronald Clark, Philip Torr; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14235-14246
We introduce Olympus, a new approach that transforms Multimodal Large Language Models (MLLMs) into a unified framework capable of handling a wide array of computer vision tasks. Utilizing a controller MLLM, Olympus delegates over 20 specialized tasks across images, videos, and 3D objects to dedicated modules. This instruction-based routing enables complex workflows through chained actions without the need for training heavy generative models. Olympus easily integrates with existing MLLMs, expanding their capabilities with comparable performance. Experimental results demonstrate that Olympus achieves an average routing accuracy of 94.75% across 20 tasks and precision of 91.82% in chained action scenarios, showcasing its effectiveness as a universal task router that can solve a diverse range of computer vision tasks.
****

HERA: Hybrid Explicit Representation for Ultra-Realistic Head Avatars
Hongrui Cai, Yuting Xiao, Xuan Wang, Jiafei Li, Yudong Guo, Yanbo Fan, Shenghua Gao, Juyong Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 260-270
We introduce a novel approach to creating ultra-realistic head avatars and rendering them in real time (\geq 30 fps at 2048 x1334 resolution). First, we propose a hybrid explicit representation that combines the advantages of two primitive based efficient rendering techniques. UV-mapped 3D mesh is utilized to capture sharp and rich textures on smooth surfaces, while 3D Gaussian Splatting is employed to represent complex geometric structures. In the pipeline of modeling an avatar, after tracking parametric models based on captured multi-view RGB videos, our goal is to simultaneously optimize the texture and opacity map of mesh, as well as a set of 3D Gaussian splats localized and rigged onto the mesh facets. Specifically, we perform \alpha-blending on the color and opacity values based on the merged and re-ordered z-buffer from the rasterization results of mesh and 3DGS. This process involves the mesh and 3DGS adaptively fitting the captured visual information to outline a high-fidelity digital avatar. To avoid artifacts caused by Gaussian splats crossing the mesh facets, we design a stable hybrid depth sorting strategy. Experiments illustrate that our modeled results exceed those of state-of-the-art approaches.
****

Circumventing Shortcuts in Audio-visual Deepfake Detection Datasets with Unsupervised Learning
Stefan Smeu, Dragos-Alexandru Boldisor, Dan Oneata, Elisabeta Oneata; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18815-18825
Good datasets are essential for developing and benchmarking any machine learning system. Their importance is even more extreme for safety critical applications such as deepfake detection - the focus of this paper. Here we reveal that two of the most widely used audio-video deepfake datasets suffer from a previously unidentified spurious feature: the leading silence. Fake videos start with a very brief moment of silence and, on the basis of this feature alone, we can separate

the real and fake samples almost perfectly. As such, previous audio-only and aud
io-video models exploit the presence of silence in the fake videos and consequen
tly perform worse when the leading silence is removed. To circumvent latching on
 such an unwanted artifact and possibly other unrevealed ones, we propose a shif
t from supervised to unsupervised learning by training models exclusively on rea
l data. We show that by aligning self-supervised audio-video representations we
remove the risk of relying on dataset-specific biases and improve robustness in
deepfake detection.
****

CoE: Chain-of-Explanation via Automatic Visual Concept Circuit Description and P
olysemanticity Quantification

Wenlong Yu, Qilong Wang, Chuang Liu, Dong Li, Qinghua Hu; Proceedings of the Com
puter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4364-4374

Explainability is a critical factor influencing the wide deployment of deep visi
on models (DVMs). Concept-based post-hoc explanation methods can provide both gl
obal and local insights into model decisions. However, current methods in this f
ield face challenges in that they are inflexible to automatically construct accu
rate and sufficient linguistic explanations for global concepts and local circui
ts. Particularly, the intrinsic polysemanticity in semantic Visual Concepts (VCs
) impedes the interpretability of concepts and DVMs, which is underestimated sev
erely. In this paper, we propose a Chain-of-Explanation (CoE) approach to addres
s these issues. Specifically, CoE automates the decoding and description of VCs
to construct global concept explanation datasets. Further, to alleviate the effe
ct of polysemanticity on model explainability, we design a concept polysemantici
ty disentanglement and filtering mechanism to distinguish the most contextually
relevant concept atoms. Besides, a Concept Polysemanticity Entropy (CPE), as a m
easure of model interpretability, is formulated to quantify the degree of concep
t uncertainty. The modeling of deterministic concepts is upgraded to uncertain c
oncept atom distributions. Ultimately, CoE automatically enables linguistic loca
l explanations of the decision-making process of DVMs by tracing the concept cir
cuit. GPT-4o and human-based experiments demonstrate the effectiveness of CPE an
d the superiority of CoE, achieving an average absolute improvement of 36% in te
rms of explainability scores.
****

Ego4o: Egocentric Human Motion Capture and Understanding from Multi-Modal Input

Jian Wang, Rishabh Dabral, Diogo Luvizon, Zhe Cao, Lingjie Liu, Thabo Beeler, Ch
ristian Theobalt; Proceedings of the Computer Vision and Pattern Recognition Con
ference (CVPR), 2025, pp. 22668-22679

This work focuses on tracking and understanding human motion using consumer wear
able devices, such as VR/AR headsets, smart glasses, cellphones, and smartwatche
s. These devices provide diverse, multi-modal sensor inputs, including egocentri
c images, and 1-3 sparse IMU sensors in varied combinations. Motion descriptions
 can also accompany these signals. The diverse input modalities and their interm
ittent availability pose challenges for consistent motion capture and understand
ing. In this work, we present Ego4o (o for omni), a new framework for simultaneo
us human motion capture and understanding from multi-modal egocentric inputs. Th
is method maintains performance with partial inputs while achieving better resul
ts when multiple modalities are combined. First, the IMU sensor inputs, the opti
onal egocentric image, and text description of human motion are encoded into the
 latent space of a motion VQ-VAE. Next, the latent vectors are sent to the VQ-VA
E decoder and optimized to track human motion. When motion descriptions are unav
ailable, the latent vectors can be input into a multi-modal LLM to generate huma
n motion descriptions, which can further enhance motion capture accuracy. Quanti
tative and qualitative evaluations demonstrate the effectiveness of our method i
n predicting accurate human motion and high-quality motion descriptions.
****

Image Over Text: Transforming Formula Recognition Evaluation with Character Dete
ction Matching

Bin Wang, Fan Wu, Linke Ouyang, Zhuangcheng Gu, Rui Zhang, Renqiu Xia, Botian Sh
i, Bo Zhang, Conghui He; Proceedings of the Computer Vision and Pattern Recognit

ion Conference (CVPR), 2025, pp. 19681-19690
Formula recognition presents significant challenges due to the complicated structure and varied notation of mathematical expressions. Despite continuous advancements in formula recognition models, the evaluation metrics employed by these models, such as BLEU and Edit Distance, still exhibit notable limitations. They overlook the fact that the same formula has diverse representations and is highly sensitive to the distribution of training data, thereby causing unfairness in formula recognition evaluation. To this end, we propose a Character Detection Matching (CDM) metric, ensuring the evaluation objectivity by designing an image-level rather than a LaTeX-level metric score. Specifically, CDM renders both the model-predicted LaTeX and the ground-truth LaTeX formulas into image-formatted formulas, then employs visual feature extraction and localization techniques for precise character-level matching, incorporating spatial position information. Such a spatially-aware and character-matching method offers a more accurate and equitable evaluation compared with previous BLEU and Edit Distance metrics that rely solely on text-based character matching. Experimentally, we evaluated various formula recognition models using CDM, BLEU, and ExpRate metrics. Their results demonstrate that the CDM aligns more closely with human evaluation standards and provides a fairer comparison across different models by eliminating discrepancies caused by diverse formula representations. Code is available at https://github.com/opendatalab/UniMERNet/tree/main/cdm.
****

FreePCA: Integrating Consistency Information across Long-short Frames in Training-free Long Video Generation via Principal Component Analysis
Jiangtong Tan, Hu Yu, Jie Huang, Jie Xiao, Feng Zhao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27979-27988
Long video generation involves generating extended videos using models trained on short videos, suffering from distribution shifts due to varying frame counts. It necessitates the use of local information from the original short frames to enhance visual and motion quality, and global information from the entire long frames to ensure appearance consistency. Existing training-free methods struggle to effectively integrate the benefits of both, as appearance and motion in videos are closely coupled, leading to inconsistency and poor quality. In this paper, we reveal that global and local information can be precisely decoupled into consistent appearance and motion intensity information by applying Principal Component Analysis (PCA), allowing for refined complementary integration of global consistency and local quality. With this insight, we propose FreePCA, a training-free long video generation paradigm based on PCA that simultaneously achieves high consistency and quality. Concretely, we decouple consistent appearance and motion intensity features by measuring cosine similarity in the principal component space. Critically, we progressively integrate these features to preserve original quality and ensure smooth transitions, while further enhancing consistency by reusing the mean statistics of the initial noise. Experiments demonstrate that FreePCA can be applied to various video diffusion models without requiring training, leading to substantial improvements. Code is available at https://github.com/JosephTiTan/FreePCA.
****

Hierarchical Adaptive Filtering Network for Text Image Specular Highlight Removal
Zhi Jiang, Jingbo Hu, Ling Zhang, Gang Fu, Chunxia Xiao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2408-2417
Despite significant advances in the field of specular highlight removal in recent years, existing methods predominantly focus on natural images, where highlights typically appear on raised or edged surfaces of objects. These highlights are often small and sparsely distributed. However, for text images such as cards and posters, the flat surfaces reflect light uniformly, resulting in large areas of highlights. Current methods struggle with these large-area highlights in text images, often producing severe visual artifacts or noticeable discrepancies between filled pixels and the original image in the central high-intensity highlight areas. To address these challenges, we propose the Hierarchical Adaptive Filteri

ng Network (HAFNet). Our approach performs filtering at both the downsampled dee p feature layer and the upsampled image reconstruction layer. By designing and applying the Adaptive Comprehensive Filtering Module (ACFM) and Adaptive Dilated Filtering Module (ADFM) at different layers, our method effectively restores se mantic information in large-area specular highlight regions and recovers detail loss at various scales. The required filtering kernels are pre-generated by a pr ediction network, allowing them to adaptively adjust according to different imag es and their semantic content, enabling robust performance across diverse scenar ios. Additionally, we utilize Unity3D to construct a comprehensive large-area h ighlight dataset featuring images with rich texts and complex textures. Experime ntal results on various datasets demonstrate that our method outperforms state-o f-the-art approaches.

****

Improving Semi-Supervised Semantic Segmentation with Sliced-Wasserstein Feature Alignment and Uniformity

Chen-Yi Lu, Kasra Derakhshandeh, Somali Chaterji; Proceedings of the Computer Vi sion and Pattern Recognition Conference (CVPR), 2025, pp. 20233-20243

Semi-supervised semantic segmentation with consistencyregularization capitalizes on unlabeled images to enhancethe accuracy of pixel-level segmentation. Current consistencylearning methods primarily rely on the consistency loss be-tween pse udo-labels and unlabeled images, neglecting the in-formation within the feature representations of the backboneencoder. Preserving maximum information in featur e embed-dings requires achieving the alignment and uniformity objec-tives, as wi dely studied. To address this, we present SWSEG,a semi-supervised semantic segme ntation algorithm that opti-mizes alignment and uniformity using the Sliced-Wass ersteinDistance (SWD), and rigorously and empirically proves thisconnection. We further resolve the computational issues as-sociated with conventional Monte Car lo-based SWD by im-plementing a Gaussian-approximated variant, which not onlymai ntains the alignment and uniformity objectives but alsoimproves training efficie ncy. We evaluate SWSEG on thePASCAL VOC 2012, Cityscapes, and ADE20K datasets, o ut-shining supervised baselines in mIoU by up to 11.8%, 8.9%,and 8.2%, respectiv ely, given an equivalent number of labeledsamples. Further, SWSEG surpasses stat e-of-the-art methodsin multiple settings across these three datasets. Our extens iveablation studies confirm the optimization of the uniformityand alignment obje ctives of the feature representations.

****

Mind the Time: Temporally-Controlled Multi-Event Video Generation

Ziyi Wu, Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Yuwei Fang, Varn ith Chordia, Igor Gilitschenski, Sergey Tulyakov; Proceedings of the Computer Vi sion and Pattern Recognition Conference (CVPR), 2025, pp. 23989-24000

Real-world videos consist of sequences of events. Generating such sequences with precise temporal control is infeasible with existing video generators that rely on a single paragraph of text as input. When tasked with generating multiple ev ents described using a single prompt, such methods often ignore some of the even ts or fail to arrange them in the correct order. To address this limitation, we present MinT, a multi-event video generator with temporal control. Our key insig ht is to bind each event to a specific period in the generated video, which allo ws the model to focus on one event at a time. To enable time-aware interactions between event captions and video tokens, we design a time-based positional encod ing method, dubbed ReRoPE. This encoding helps to guide the cross-attention oper ation. By fine-tuning a pre-trained video diffusion transformer on temporally gr ounded data, our approach produces coherent videos with smoothly connected event s. For the first time in the literature, our model offers control over the timin g of events in generated videos. Extensive experiments demonstrate that MinT out performs existing commercial and open-source models by a large margin.

****

Learning Extremely High Density Crowds as Active Matters

Feixiang He, Jiangbei Yue, Jialin Zhu, Armin Seyfried, Dan Casas, Julien Pettré, He Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 540-550

Video-based high-density crowd analysis and prediction has been a long-standing topic in computer vision. It is notoriously difficult due to, but not limited to , the lack of high-quality data and complex crowd dynamics. Consequently, it has been relatively under studied. In this paper, we propose a new approach that aims to learn from in-the-wild videos, often with low quality where it is difficult to track individuals or count heads. The key novelty is a new physics prior to model crowd dynamics. We model high-density crowds as active matter, a continuum with active particles subject to stochastic forces, named `crowd material'. Our physics model is combined with neural networks, resulting in a neural stochastic differential equation system which can mimic the complex crowd dynamics. Due to the lack of similar research, we adapt a range of existing methods which are close to ours, and other possible alternatives for comparison. Through exhaustive evaluation, we show our model outperforms existing methods in analyzing and forecasting extremely high-density crowds. Furthermore, since our model is a continuous-time physics model, it can be used for simulation and analysis, providing strong interpretability. This is categorically different from most deep learning methods, which are discrete-time models and black-boxes.

****

Audio-Visual Semantic Graph Network for Audio-Visual Event Localization
Liang Liu, Shuaiyong Li, Yongqiang Zhu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23957-23966
Audio-visual event localization (AVEL) aims to identify both the category and temporal boundaries of events that are both audible and visible in unconstrained videos. However, the inherent semantic gap between heterogeneous modalities often leads to semantic inconsistency. In this paper, we propose a novel Audio-Visual Semantic Graph Network (AVSGN) to facilitate cross-modal alignment and cross-temporal interaction. Unlike previous approaches (e.g., audio-guided, visual-guided, or both), we introduce shared semantic textual labels to bridge the semantic gap between audio and visual modalities. Specifically, we present a cross-modal semantic alignment (CMSA) module to explore the complementary relationships across heterogeneous modalities (i.e., visual, audio, and text), promoting the convergence of multimodal distributions into a unified semantic space. Additionally, in order to capture cross-temporal dependencies sufficiently, we devise a cross-modal graph interaction (CMGI) module which disentangles complicated interactions across modalities into three complementary subgraphs. Extensive experiments on the AVE dataset comprehensively demonstrate the superiority and effectiveness of the proposed model in both fully- and weakly-supervised AVE settings.

****

3D-Mem: 3D Scene Memory for Embodied Exploration and Reasoning
Yuncong Yang, Han Yang, Jiachen Zhou, Peihao Chen, Hongxin Zhang, Yilun Du, Chuang Gan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17294-17303
Constructing compact and informative 3D scene representations is essential for effective embodied exploration and reasoning, especially in complex environments over extended periods. Existing representations, such as object-centric 3D scene graphs, oversimplify spatial relationships by modeling scenes as isolated objects with restrictive textual relationships, making it difficult to address queries requiring nuanced spatial understanding. Moreover, these representations lack natural mechanisms for active exploration and memory management, hindering their application to lifelong autonomy. In this work, we propose 3D-Mem, a novel 3D scene memory framework for embodied agents. 3D-Mem employs informative multi-view images, termed Memory Snapshots, to capture rich visual information of explored regions. It further integrates frontier-based exploration by introducing Frontier Snapshots--glimpses of unexplored areas--enabling agents to make decisions by considering both known and potential new information. To support lifelong memory in active exploration settings, we present an incremental construction pipeline for 3D-Mem, as well as a memory retrieval technique for memory management. Experimental results on three benchmarks demonstrate that 3D-Mem significantly enhances agents' exploration and reasoning capabilities in 3D environments, highlighting its potential for advancing applications in embodied AI.

****
EchoMimicV2: Towards Striking, Simplified, and Semi-Body Human Animation
Rang Meng, Xingyu Zhang, Yuming Li, Chenguang Ma; Proceedings of the Computer Vi sion and Pattern Recognition Conference (CVPR), 2025, pp. 5489-5498
Recent work on human animation usually involves audio, pose, or movement maps co nditions, thereby achieves vivid animation quality. However, these methods often face practical challenges due to extra control conditions, cumbersome condition injection modules, or limitation to head region driving. Hence, we ask if it is possible to achieve striking half-body human animation while simplifying unnece ssary conditions. To this end, we propose a half-body human animation method, du bbed EchoMimicV2, that leverages a novel Audio-Pose Dynamic Harmonization strate gy, including Pose Sampling and Audio Diffusion, to enhance half-body details, f acial and gestural expressiveness, and meanwhile reduce conditions redundancy. T o compensate for the scarcity of half-body data, we utilize Head Partial Attenti on to seamlessly accommodate headshot data into our training framework, which ca n be omitted during inference, providing a free lunch for animation. Furthermore , we design the Phase-specific Denoising Loss to guide motion, detail, and low-l evel quality for animation in specific phases, respectively. Besides, we also pr esent a novel benchmark for evaluating the effectiveness of half-body human anim ation. Extensive experiments and analyses demonstrate that EchoMimicV2 surpasses existing methods in both quantitative and qualitative evaluations.
****
Navigation World Models
Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, Yann LeCun; Proceedings of th e Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15791-158 01
Navigation is a fundamental skill of agents with visual-motor capabilities. We i ntroduce a Navigation World Model (NWM), a controllable video generation model t hat predicts future visual observations based on past observations and navigatio n actions. To capture complex environment dynamics, NWM employs a Conditional Di ffusion Transformer (CDiT), trained on a diverse collection of egocentric videos of both human and robotic agents, and scaled up to 1 billion parameters. In fam iliar environments, NWM can plan navigation trajectories by simulating them and evaluating whether they achieve the desired goal. Unlike supervised navigation p olicies with fixed behavior, NWM can dynamically incorporate constraints during planning. Experiments demonstrate its effectiveness in planning trajectories fro m scratch or by ranking trajectories sampled from an external policy. Furthermor e, NWM leverages its learned visual priors to imagine trajectories in unfamiliar environments from a single input image, making it a flexible and powerful tool for next-generation navigation systems.
****
Video Motion Transfer with Diffusion Transformers
Alexander Pondaven, Aliaksandr Siarohin, Sergey Tulyakov, Philip Torr, Fabio Piz zati; Proceedings of the Computer Vision and Pattern Recognition Conference (CVP R), 2025, pp. 22911-22921
We propose DiTFlow, a method for transferring the motion of a reference video to a newly synthesized one, designed specifically for Diffusion Transformers (DiT) . We first process the reference video with a pre-trained DiT to analyze cross-f rame attention maps and extract a patch-wise motion signal called the Attention Motion Flow (AMF). We guide the latent denoising process in an optimization-base d, training-free, manner by optimizing latents with our AMF loss to generate vid eos reproducing the motion of the reference one. We also apply our optimization strategy to transformer positional embeddings, granting us a boost in zero-shot motion transfer capabilities. We evaluate DiTFlow against recently published met hods, outperforming all across multiple metrics and human evaluation.
****
Gaussian Splatting for Efficient Satellite Image Photogrammetry
Luca Savant Aira, Gabriele Facciolo, Thibaud Ehret; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5959-5969
Recently, Gaussian splatting has emerged as a strong alternative to NeRF, demons

trating impressive 3D modeling capabilities while requiring only a fraction of t
he training and rendering time. In this paper, we show how the standard Gaussian
 splatting framework can be adapted for remote sensing, retaining its high effic
iency. This enables us to achieve state-of-the-art performance in just a few min
utes, compared to the day-long optimization required by the best-performing NeRF
-based Earth observation methods. The proposed framework incorporates remote-sen
sing improvements from EO-NeRF, such as radiometric correction and shadow modeli
ng, while introducing novel components, including sparsity, view consistency, an
d opacity regularizations.
****

Unified Reconstruction of Static and Dynamic Scenes from Events
Qiyao Gao, Peiqi Duan, Hanyue Lou, Minggui Teng, Ziqi Cai, Xu Chen, Boxin Shi; P
roceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 202
5, pp. 27914-27923
This paper addresses the challenge that current event-based video reconstruction
 methods cannot produce static background information. Recent research has uncov
ered the potential of event cameras in capturing static scenes. Nonetheless, ima
ge quality deteriorates due to noise interference and detail loss, failing to pr
ovide reliable background information. We propose a two-stage reconstruction str
ategy to address these challenges and reconstruct static scene images comparable
 to frame cameras. Building on this, we introduce the URSEE framework, the first
 unified framework designed for reconstructing motion videos with static backgro
unds. This framework includes a parallel channel that can simultaneously process
 static and dynamic events, and a network module designed to reconstruct videos
encompassing both static and dynamic scenes in an end-to-end manner. We also col
lect a real-captured dataset for static reconstruction, containing both indoor a
nd outdoor scenes. Comparison results indicate that the proposed approach achiev
es state-of-the-art reconstruction results on both synthetic and real data.
****

Automatic Spectral Calibration of Hyperspectral Images: Method, Dataset and Benc
hmark
Zhuoran Du, Shaodi You, Cheng Cheng, Shikui Wei; Proceedings of the Computer Vis
ion and Pattern Recognition Conference (CVPR), 2025, pp. 28081-28090
Hyperspectral images (HSI) densely sample the world in both the space and freque
ncy domains and, therefore, are more distinctive than RGB images. Usually, HSI n
eeds to be calibrated to minimize the impact of various illumination conditions.
 The traditional way to calibrate HSI utilizes a physical reference, which invol
ves manual operations, occlusions, and/or limits camera mobility. These limitati
ons inspire this paper to automatically calibrate HSIs using a learning-based me
thod. Towards this goal, a large-scale HSI calibration dataset, which has 765 hi
gh-quality HSI pairs covering diversified natural scenes and illuminations, is c
reated. The dataset is further expanded to 7650 pairs by combining with 10 diffe
rent physically measured illuminations. A spectral illumination transformer (SIT
) together with an illumination attention module is proposed. Extensive benchmar
ks demonstrate the SoTA performance of the proposed SIT. The benchmarks also ind
icate that low-light conditions are more challenging than normal conditions. The
 dataset and codes are available online: https://github.com/duranze/Automatic-sp
ectral-calibration-of-HSI.
****

Conformal Prediction and MLLM aided Uncertainty Quantification in Scene Graph Ge
neration
Sayak Nag, Udita Ghosh, Calvin-Khang Ta, Sarosij Bose, Jiachen Li, Amit K. Roy-C
howdhury; Proceedings of the Computer Vision and Pattern Recognition Conference
(CVPR), 2025, pp. 11676-11686
Scene Graph Generation (SGG) aims to represent visual scenes by identifying obje
cts and their pairwise relationships, providing a structured understanding of im
age content. However, inherent challenges like long-tailed class distributions a
nd prediction variability necessitate uncertainty quantification in SGG for its
practical viability. In this paper, we introduce a novel Conformal Prediction ba
sed framework, adaptive to any existing SGG method, for quantifying their predic

tive uncertainty by constructing well-calibrated prediction sets over their gene rated scene graphs. These scene graph prediction sets are designed to achieve st atistically rigorous coverage guarantees under exchangeability assumptions. Addi tionally, to ensure the prediction sets contain the most practically interpretab le scene graphs, we propose an effective MLLM-based post-processing strategy for selecting the most visually and semantically plausible scene graphs within each set. We show that our proposed approach can produce diverse possible scene grap hs from an image, assess the reliability of SGG methods, and improve overall SGG performance.
****

Point-to-Region Loss for Semi-Supervised Point-Based Crowd Counting
Wei Lin, Chenyang Zhao, Antoni B. Chan; Proceedings of the Computer Vision and P attern Recognition Conference (CVPR), 2025, pp. 29363-29373
Point detection has been developed to locate pedestrians in crowded scenes by tr aining a counter through a point-to-point (P2P) supervision scheme. Despite its excellent localization and counting performance, training a point-based counter still faces challenges concerning annotation labor: hundreds to thousands of poi nts are required to annotate a single sample capturing a dense crowd. In this pa per, we integrate point-based methods into a semi-supervised counting framework based on pseudo-labeling, enabling the training of a counter with only a few ann otated samples supplemented by a large volume of pseudo-labeled data. However, d uring implementation, the training encounters issues as the confidence for pseud o-labels fails to be propagated to background pixels via the P2P. To tackle this challenge, we devise a point-specific activation map (PSAM) to visually interpr et the phenomena occurring during the ill-posed training. Observations from the PSAM suggest that the feature map is excessively activated by the loss for unlab eled data, causing the decoder to misinterpret these over-activations as pedestr ians. To mitigate this issue, we propose a point-to-region (P2R) scheme to subst itute P2P, which segments out local regions rather than detects a point correspo nding to a pedestrian for supervision. Consequently, pixels in the local region can share the same confidence with the corresponding pseudo points. Experimental results in both semi-supervised counting and unsupervised domain adaptation hig hlight the advantages of our method, illustrating P2R can resolve issues identif ied in PSAM. The code is available at https://github.com/Elin24/P2RLoss.
****

Reconstructing Close Human Interaction with Appearance and Proxemics Reasoning
Buzhen Huang, Chen Li, Chongyang Xu, Dongyue Lu, Jinnan Chen, Yangang Wang, Gim Hee Lee; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17475-17485
Due to visual ambiguities and inter-person occlusions, existing human pose estim ation methods cannot recover plausible close interactions from in-the-wild video s. Even state-of-the-art large foundation models (e.g., SAM) cannot accurately d istinguish human semantics in such challenging scenarios. In this work, we find that human appearance can provide a straightforward cue to address these obstacl es. Based on this observation, we propose a dual-branch optimization framework t o reconstruct accurate interactive motions with plausible body contacts constrai ned by human appearances, social proxemics, and physical laws. Specifically, we first train a diffusion model to learn the human proxemic behavior and pose prio r knowledge. The trained network and two optimizable tensors are then incorporat ed into a dual-branch optimization framework to reconstruct human motions and ap pearances. Several constraints based on 3D Gaussians, 2D keypoints, and mesh pen etrations are also designed to assist the optimization. With the proxemics prior and diverse constraints, our method is capable of estimating accurate interacti ons from in-the-wild videos captured in complex environments. We further build a dataset with pseudo ground-truth interaction annotations, which may promote fut ure research on pose estimation and human behavior understanding. Experimental r esults on several benchmarks demonstrate that our method outperforms existing ap proaches. The code and data are available at https://www.buzhenhuang.com/works/C loseApp.html.
****

Towards Improved Text-Aligned Codebook Learning: Multi-Hierarchical Codebook-Text Alignment with Long Text
Guotao Liang, Baoquan Zhang, Zhiyuan Wen, Junteng Zhao, Yunming Ye, Kola Ye, Yao He; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4060-4069
Image quantization is a crucial technique in image generation, aimed at learning a codebook that encodes an image into a discrete token sequence. Recent advancements have seen researchers exploring learning multi-modal codebook (i.e., text-aligned codebook) by utilizing image caption semantics, aiming to enhance codebook performance in cross-modal tasks. However, existing image-text paired datasets exhibit a notable flaw in that the text descriptions tend to be overly concise, failing to adequately describe the images and provide sufficient semantic knowledge, resulting in limited alignment of text and codebook at a fine-grained level.In this paper, we propose a novel Text-Augmented Codebook Learning framework, named TA-VQ, which generates longer text for each image using the visual-language model for improved text-aligned codebook learning.However, the long text presents two key challenges: how to encode text and how to align codebook and text. To tackle two challenges, we propose to split the long text into multiple granularities for encoding, i.e., word, phrase, and sentence, so that the long text can be fully encoded without losing any key semantic knowledge. Following this, a hierarchical encoder and novel sampling-based alignment strategy are designed to achieve fine-grained codebook-text alignment. Additionally, our method can be seamlessly integrated into existing VQ models. Extensive experiments in reconstruction and various downstream tasks demonstrate its effectiveness compared to previous state-of-the-art approaches.
****
Parallel Sequence Modeling via Generalized Spatial Propagation Network
Hongjun Wang, Wonmin Byeon, Jiarui Xu, Jinwei Gu, Ka Chun Cheung, Xiaolong Wang, Kai Han, Jan Kautz, Sifei Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4473-4483
We present the Generalized Spatial Propagation Network (GSPN), a new attention mechanism optimized for vision tasks that inherently captures 2D spatial structures. Existing attention models, including transformers, linear attention, and state-space models like Mamba, process multi-dimensional data as 1D sequences, compromising spatial coherence and efficiency. GSPN overcomes these limitations by directly operating on spatially coherent image data and forming dense pairwise connections through a unique line-scan approach. Central to GSPN is the Stability-Context Condition, which ensures stable, context-aware propagation across 2D sequences and reduces the effective sequence length to \sqrt N , significantly enhancing computational efficiency. With learnable, input-dependent weights and no reliance on positional embeddings, GSPN achieves superior spatial fidelity and state-of-the-art performance in vision tasks, including ImageNet classification, class-guided image generation, and text-to-image generation. Notably, GSPN accelerates SD-XL with softmax-attention by over 84xwhen generating 16K images.
****
Scenario Dreamer: Vectorized Latent Diffusion for Generating Driving Simulation Environments
Luke Rowe, Roger Girgis, Anthony Gosselin, Liam Paull, Christopher Pal, Felix Heide; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17207-17218
We introduce Scenario Dreamer, a fully data-driven generative simulator for autonomous vehicle planning that generates both the initial traffic scene--comprising a lane graph and agent bounding boxes--and closed-loop agent behaviours. Existing methods for generating driving simulation environments encode the initial traffic scene as a rasterized image and, as such, require parameter-heavy networks that perform unnecessary computation due to many empty pixels in the rasterized scene. Moreover, we find that existing methods that employ rule-based agent behaviours lack diversity and realism. Scenario Dreamer instead employs a novel vectorized latent diffusion model for initial scene generation that directly operates on the vectorized scene elements and an autoregressive Transformer for data-d

riven agent behaviour simulation. Scenario Dreamer additionally supports scene e
xtrapolation via diffusion inpainting, enabling the generation of unbounded simu
lation environments. Extensive experiments show that Scenario Dreamer outperform
s existing generative simulators in realism and efficiency: the vectorized scene
-generation base model achieves superior generation quality with around 2x fewer
 parameters, 6x lower generation latency, and 10x fewer GPU training hours compa
red to the strongest baseline. We confirm its practical utility by showing that
reinforcement learning planning agents are more challenged in Scenario Dreamer e
nvironments than traditional non-generative simulation environments, especially
on long and adversarial driving environments.
****

Poly-Autoregressive Prediction for Modeling Interactions
Neerja Thakkar, Tara Sadjadpour, Jathushan Rajasegeran, Shiry Ginosar, Jitendra
Malik; Proceedings of the Computer Vision and Pattern Recognition Conference (CV
PR), 2025, pp. 12402-12412
We introduce a simple framework for predicting the behavior of an agent in multi
-agent settings. In contrast to autoregressive (AR) tasks, such as language proc
essing, our focus is on scenarios with multiple agents whose interactions are sh
aped by physical constraints and internal motivations. To this end, we propose P
oly-Autoregressive (PAR) modeling, which forecasts an ego agent's future behavio
r by reasoning about the ego agent's state history and the past and current stat
es of other interacting agents. At its core, PAR represents the behavior of all
agents as a sequence of tokens, each representing an agent's state at a specific
 timestep. With minimal data pre-processing changes, we show that PAR can be app
lied to three different problems: human action forecasting in social situations,
 trajectory prediction for autonomous vehicles, and object pose forecasting duri
ng hand-object interaction. Using a small proof-of-concept transformer backbone,
 PAR outperforms AR across these three scenarios.
****

NADER: Neural Architecture Design via Multi-Agent Collaboration
Zekang Yang, Wang Zeng, Sheng Jin, Chen Qian, Ping Luo, Wentao Liu; Proceedings
of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4452
-4461
Designing effective neural architectures poses a significant challenge in deep l
earning. While Neural Architecture Search (NAS) automates the search for optimal
 architectures, existing methods are often constrained by predetermined search s
paces and may miss critical neural architectures. In this paper, we introduce NA
DER (Neural Architecture Design via multi-agEnt collaboRation), a novel framewor
k that formulates neural architecture design (NAD) as a LLM-based multi-agent co
llaboration problem. NADER employs a team of specialized agents to enhance a bas
e architecture through iterative modification. Current LLM-based NAD methods typ
ically operate independently, lacking the ability to learn from past experiences
, which results in repeated mistakes and inefficient exploration. To address thi
s issue, we propose the Reflector, which effectively learns from immediate feedb
ack and long-term experiences. Additionally, unlike previous LLM-based methods t
hat use code to represent neural architectures, we utilize a graph-based represe
ntation. This approach allows agents to focus on design aspects without being di
stracted by coding. We demonstrate the effectiveness of NADER in discovering hig
h-performing architectures beyond predetermined search spaces through extensive
experiments on benchmark tasks, showcasing its advantages over state-of-the-art
methods.
****

Move-in-2D: 2D-Conditioned Human Motion Generation
Hsin-Ping Huang, Yang Zhou, Jui-Hsien Wang, Difan Liu, Feng Liu, Ming-Hsuan Yang
, Zhan Xu; Proceedings of the Computer Vision and Pattern Recognition Conference
 (CVPR), 2025, pp. 22766-22775
Generating realistic human videos remains a challenging task, with the most effe
ctive methods currently relying on a human motion sequence as a control signal.
Existing approaches often use existing motion extracted from other videos, which
 restricts applications to specific motion types and global scene matching. We p

ropose Move-in-2D, a novel approach to generate human motion sequences condition ed on a scene image, allowing for diverse motion that adapts to different scenes . Our approach utilizes a diffusion model that accepts both a scene image and te xt prompt as inputs, producing a motion sequence tailored to the scene. To train this model, we collect a large-scale video dataset featuring single-human activ ities, annotating each video with the corresponding human motion as the target o utput. Experiments demonstrate that our method effectively predicts human motion that aligns with the scene image after projection. Furthermore, we show that th e generated motion sequence improves human motion quality in video synthesis tas ks.

****

PoseBH: Prototypical Multi-Dataset Training Beyond Human Pose Estimation
Uyoung Jeong, Jonathan Freer, Seungryul Baek, Hyung Jin Chang, Kwang In Kim; Pro ceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12278-12288

We study multi-dataset training (MDT) for pose estimation, where skeletal hetero geneity presents a unique challenge that existing methods have yet to address. I n traditional domains, e.g. regression and classification, MDT typically relies on dataset merging or multi-head supervision. However, the diversity of skeleton types and limited cross-dataset supervision complicate integration in pose esti mation. To address these challenges, we introduce PoseBH, a new MDT framework th at tackles keypoint heterogeneity and limited supervision through two key techni ques. First, we propose nonparametric keypoint prototypes that learn within a un ified embedding space, enabling seamless integration across skeleton types. Seco nd, we develop a cross-type self-supervision mechanism that aligns keypoint pred ictions with keypoint embedding prototypes, providing supervision without relyin g on teacher-student models or additional augmentations. PoseBH substantially im proves generalization across whole-body and animal pose datasets, including COCO -WholeBody, AP-10K, and APT-36K, while preserving performance on standard human pose benchmarks (COCO, MPII, and AIC). Furthermore, our learned keypoint embeddi ngs transfer effectively to hand shape estimation (InterHand2.6M) and human body shape estimation (3DPW). The code for PoseBH is available at: `https://github.c om/uyoung-jeong/PoseBH`.

****

MATCHA: Towards Matching Anything
Fei Xue, Sven Elflein, Laura Leal-Taixé, Qunjie Zhou; Proceedings of the Compute r Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27081-27091

Establishing correspondences across images is a fundamental challenge in compute r vision, underpinning tasks like Structure-from-Motion, image editing, and poin t tracking. Traditional methods are often specialized for specific correspondenc e types, geometric, semantic, or temporal, whereas humans naturally identify ali gnments across these domains. Inspired by this flexibility, we propose MATCHA, a unified feature model designed to "rule them all", establishing robust correspo ndences across diverse matching tasks. Building on insights that diffusion model features can encode multiple correspondence types, MATCHA augments this capacit y by dynamically fusing high-level semantic and low-level geometric features thr ough an attention-based module, creating expressive, versatile, and robust featu res. Additionally, MATCHA integrates object-level features from DINOv2 to furthe r boost generalization, enabling a single feature capable of matching anything. Extensive experiments validate that MATCHA consistently surpasses state-of-the-a rt methods across geometric, semantic, and temporal matching tasks, setting a ne w foundation for a unified approach for the fundamental correspondence problem i n computer vision. To the best of our knowledge, MATCHA is the first approach th at is able to effectively tackle diverse matching tasks with a single unified fe ature. Project page: https://github.com/feixue94/matcha.

****

CTRL-D: Controllable Dynamic 3D Scene Editing with Personalized 2D Diffusion
Kai He, Chin-Hsuan Wu, Igor Gilitschenski; Proceedings of the Computer Vision an d Pattern Recognition Conference (CVPR), 2025, pp. 26630-26640

Achieving controllable and consistent editing in dynamic 3D scenes remains a sig

nificant challenge. Previous work is largely constrained by its editing backbone
s, resulting in inconsistent edits and limited controllability. We propose to ad
dress this challenge using personalized diffusion models. In our work, we introd
uce a novel framework that first fine-tunes the InstructPix2Pix model, followed
by a two-stage optimization of the scene based on deformable 3D Gaussians. The f
ine-tuning enables the model to "learn" the editing ability from a single edited
 reference image. This reduces the more complex task of dynamic scene editing to
 the more simple 2D image editing problem. By directly learning editing regions
and styles from the reference, our approach enables consistent and precise local
 edits without the need for tracking desired editing regions, effectively addres
sing key challenges in dynamic scene editing. Then, our two-stage optimization p
rogressively edits the trained dynamic scene, using a designed edited image buff
er to accelerate convergence and improve temporal consistency. Compared to state
-of-the-art methods, our approach offers more flexible and controllable local sc
ene editing, achieving high-quality and consistent results.
****

Separation of Powers: On Segregating Knowledge from Observation in LLM-enabled K
nowledge-based Visual Question Answering

Zhen Yang, Zhuo Tao, Qi Chen, Liang Li, Yuankai Qi, Anton van den Hengel, Qingmi
ng Huang; Proceedings of the Computer Vision and Pattern Recognition Conference
(CVPR), 2025, pp. 24753-24762

Knowledge-based visual question answering (KBVQA) separates image interpretation
 and knowledge retrieval into separate processes, motivated in part by the fact
that they are very different tasks. In this paper, we transform the KBVQA into l
inguistic question-answering tasks so that we can leverage the rich world knowle
dge and strong reasoning abilities of Large Language Models (LLMs). The caption-
then-question approach to KBVQA has been effective but relies on the captioning
method to describe the detail required to answer every possible question. We pro
pose instead a Question-Aware Captioner (QACap), which uses the question as guid
ance to extract correlated visual information from the image and generate a ques
tion-related caption. To train such a model, we utilize GPT-4 to build a corresp
onding high-quality question-aware caption dataset on top of existing KBVQA data
sets. Extensive experiments demonstrate that our QACap model and dataset signifi
cantly improve KBVQA performance. Our method, QACap, achieves 68.2% accuracy on
the OKVQA validation set, 73.4% on the direct-answer part of the A-OKVQA validat
ion set, and 74.8% on the multiple-choice part, all setting new SOTA benchmarks.
****

Decision SpikeFormer: Spike-Driven Transformer for Decision Making

Wei Huang, Qinying Gu, Nanyang Ye; Proceedings of the Computer Vision and Patter
n Recognition Conference (CVPR), 2025, pp. 19241-19250

Offline reinforcement learning (RL) enables policy training solely on pre-collec
ted data, avoiding direct environment interaction--a crucial benefit for energy-
constrained embodied AI applications. Although Artificial Neural Networks (ANN)-
based methods perform well in offline RL, their high computational and energy de
mands motivate exploration of more efficient alternatives. Spiking Neural Networ
ks (SNNs) show promise for such tasks, given their low power consumption. In thi
s work, we introduce DSFormer, the first spike-driven transformer model designed
 to tackle offline RL via sequence modeling. Unlike existing SNN transformers fo
cused on spatial dimensions for vision tasks, we develop Temporal Spiking Self-A
ttention (TSSA) and Positional Spiking Self-Attention (PSSA) in DSFormer to capt
ure the temporal and positional dependencies essential for sequence modeling in
RL. Additionally, we propose Progressive Threshold-dependent Batch Normalization
 (PTBN), which combines the benefits of LayerNorm and BatchNorm to preserve temp
oral dependencies while maintaining the spiking nature of SNNs. Comprehensive re
sults in the D4RL benchmark show DSFormer's superiority over both SNN and ANN co
unterparts, achieving 78.4% energy savings, highlighting DSFormer's advantages n
ot only in energy efficiency but also in competitive performance. Code and model
s are public at \href https://wei-nijuan.github.io/DecisionSpikeFormer/ project
 page .
****

SF2T: Self-supervised Fragment Finetuning of Video-LLMs for Fine-Grained Understanding

Yangliu Hu, Zikai Song, Na Feng, Yawei Luo, Junqing Yu, Yi-Ping Phoebe Chen, Wei Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29108-29117

Video-based Large Language Models (Video-LLMs) have witnessed substantial advancements in recent years, propelled by the advancement in multi-modal LLMs. Although these models have demonstrated proficiency in providing the overall description of videos, they struggle with fine-grained understanding, particularly in aspects such as visual dynamics and video details inquiries. To tackle these shortcomings, we find that fine-tuning Video-LLMs on self-supervised fragment tasks, greatly improve their fine-grained video understanding abilities. Hence we propose two key contributions:(1) Self-Supervised Fragment Fine-Tuning (SF^2T), a novel effortless fine-tuning method, employs the rich inherent characteristics of videos for training, while unlocking more fine-grained understanding ability of Video-LLMs. Moreover, it relieves researchers from labor-intensive annotations and smartly circumvents the limitations of natural language, which often fails to capture the complex spatiotemporal variations in videos;(2) A novel benchmark dataset, namely FineVidBench, for rigorously assessing Video-LLMs' performance at both the scene and fragment levels, offering a comprehensive evaluation of their capabilities.We assessed multiple models and validated the effectiveness of SF^2T on them. Experimental results reveal that our approach improves their ability to capture and interpret spatiotemporal details.

****

Theory-Inspired Deep Multi-View Multi-Label Learning with Incomplete Views and Noisy Labels

Quanjiang Li, Tingjin Luo, Jiahui Liao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20706-20715

Incomplete features and label noise in multi-view multi-label data significantly undermine the reliability and performance, motivating researchers to explore the mechanism of representation and information recovery. However, learning for such dual deficiencies is crucial but rarely studied. In this paper, we propose a theory-inspired Deep Multi-View Multi-Label Learning method with Incomplete Views and Noisy Labels named DMMIvNL to address these problems. Specifically, to promote the synthesis of task-relevant shared information and preserve the distinctiveness of individual features from limited views, we have developed a feature extraction modular based on the information bottleneck theory, and formulated its theoretical upper bound into its objective. Meanwhile, we theoretically prove that minimizing the volume of the transition matrix ensures the statistical consistency with classifier training. Besides, a cycle-consistency estimation principle is proposed in the volume minimization network to improve the recognition stability of multi-label noise. Moreover, leveraging inherent real semantics information and label correlations are employed as model regularization to reduce the risk of excessive noise fitting. Finally, extensive experimental results validate the effectiveness and robustness of our DMMIvNL.

****

Fitted Neural Lossless Image Compression

Zhe Zhang, Zhenzhong Chen, Shan Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23249-23258

Neural lossless image compression methods have recently achieved impressive compression ratios by fitting neural networks to represent data distributions of large datasets. However, these methods often require complex networks to capture intricate data distributions effectively, resulting in high decoding complexity. In this paper, we present a novel approach named Fitted Neural Lossless Image Compression (FNLIC) that enhances efficiency through a two-phase fitting process. For each image, a latent variable model is overfitted to optimize the representation of the individual image's probability distribution, which is inherently simpler than the distribution of an entire dataset and requires less complex neural networks. Additionally, we pre-fit a lightweight autoregressive model on a comprehensive dataset to learn a beneficial prior for overfitted models. To improve c

oordination between the pre-fitting and overfitting phases, we introduce indepen dent fitting for the pre-fitter and the adaptive prior transformation for the ov erfitted model. Extensive experimental results on high-resolution datasets show that FNLIC achieves competitive compression ratios compared to both traditional and neural lossless image compression methods, with decoding complexity signific antly lower than other neural methods of similar performance. The code is at htt ps://github.com/ZZ022/FNLIC.
****

Fortifying Federated Learning Towards Trustworthiness via Auditable Data Valuati on and Verifiable Client Contribution

K Naveen Kumar, Ranjeet Ranjan Jha, C Krishna Mohan, Ravindra Babu Tallamraju; P roceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 202 5, pp. 4999-5009

Ensuring auditability and verifiability in FL is both challenging and essential to guarantee that local data remains untampered and client updates are trustwort hy. Recent FL frameworks assess client contributions through a trusted central s erver using various client selection and aggregation techniques. However, relian ce on a central server can create a single point of failure, making it vulnerabl e to privacy-centric attacks and limiting its ability to audit and verify client -side data contributions due to restricted access. In addition, data quality and fairness evaluations are often inadequate, failing to distinguish between high- impact contributions and those from low-quality or poisoned data. To address the se challenges, we propose Federated Auditable and Verifiable Data valuation (FAV D), a privacy-preserving method that ensures auditability and verifiability of c lient contributions through data valuation, independent of any central authority or predefined training algorithm. FAVD utilizes shared local data density funct ions to construct a global density function, aligning data contributions and fac ilitating effective valuation prior to local model training. This proactive appr oach improves transparency in data valuation and ensures that only benign update s are generated, even in the presence of malicious data. Further, to mitigate pr ivacy risks associated with sharing data density functions, we add Gaussian nois e to each client's local density function before sharing it with the server. We theoretically demonstrate the convergence, auditability, and verifiability of FA VD, along with its resilience against data poisoning threats. Our experiments on five diverse benchmarks, including three medical datasets, show that FAVD achie ves significant performance gains, accurate data valuation, and fair client cont ributions under threat, highlighting its reliability as a trustworthy FL approac h.
****

EMOE: Modality-Specific Enhanced Dynamic Emotion Experts

Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, Mang Ye; Proceedings of the C omputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14314-14324

Multimodal Emotion Recognition (MER) aims to predict human emotions by leveragin g multiple modalities, such as vision, acoustics, and language. However, due to the heterogeneity of these modalities, MER faces two key challenges: modality ba lance dilemma and modality specialization disappearance. Existing methods often overlook the varying importance of modalities across samples in tackling the mod ality balance dilemma. Moreover, mainstream decoupling methods, while preserving modality-specific information, often neglect the predictive capability of unimo dal data. To address these, we propose a novel model, Modality-Specific Enhanced Dynamic Emotion Experts (EMOE), consisting of: (1) Mixture of Modality Experts for dynamically adjusting modality importance based on sample features, and (2) Unimodal Distillation to retain single-modality predictive ability within fused features. EMOE enables adaptive fusion by learning a unique modality weight dist ribution for each sample, enhancing multimodal predictions with single-modality predictions to balance invariant and specific features in emotion recognition. E xperimental results on benchmark datasets show that EMOE achieves superior or co mparable performance to state-of-the-art methods. Additionally, we extend EMOE t o Multimodal Intent Recognition (MIR), further demonstrating its effectiveness a nd versatility.

****
JarvisIR: Elevating Autonomous Driving Perception with Intelligent Image Restoration
Yunlong Lin, Zixu Lin, Haoyu Chen, Panwang Pan, Chenxin Li, Sixiang Chen, Kairun Wen, Yeying Jin, Wenbo Li, Xinghao Ding; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22369-22380
Vision-centric perception systems often struggle with unpredictable and coupled weather degradations in the wild. Current solutions are often limited, as they either depend on specific degradation priors or suffer from significant domain gaps. To enable robust and autonomous operation in real-world conditions, we propose JarvisIR, a VLM-powered agent that leverages the VLM (e.g., Llava-Llama3) as a controller to manage multiple expert restoration models. To further enhance system robustness, reduce hallucinations, and improve generalizability in real-world adverse weather, JarvisIR employs a novel two-stage framework consisting of supervised fine-tuning and human feedback alignment. Specifically, to address the lack of paired data in real-world scenarios, the human feedback alignment enables the VLM to be fine-tuned effectively on large-scale real-world data in an unsupervised manner. To support the training and evaluation of JarvisIR, we introduce CleanBench, a comprehensive dataset consisting of high-quality and large-scale instruction-responses pairs, including 150K synthetic entries and 80K real entries. Extensive experiments demonstrate that JarvisIR exhibits superior decision-making and restoration capabilities. Compared with existing methods, it achieves a 50% improvement in the average of all perception metrics on CleanBench-Real.
****
UniPre3D: Unified Pre-training of 3D Point Cloud Models with Cross-Modal Gaussian Splatting
Ziyi Wang, Yanran Zhang, Jie Zhou, Jiwen Lu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1319-1329
The scale diversity of point cloud data presents significant challenges in developing unified representation learning techniques for 3D vision. Currently, there are few unified 3D models, and no existing pre-training method is equally effective for both object- and scene-level point clouds. In this paper, we introduce UniPre3D, the first unified pre-training method that can be seamlessly applied to point clouds of any scale and 3D models of any architecture. Our approach predicts Gaussian primitives as the pre-training task and employs differentiable Gaussian splatting to render images, enabling precise pixel-level supervision and end-to-end optimization. To further regulate the complexity of the pre-training task and direct the model's focus toward geometric structures, we integrate 2D features from pre-trained image models to incorporate well-established texture knowledge. We validate the universal effectiveness of our proposed method through extensive experiments across a variety of object- and scene-level tasks, using diverse point cloud models as backbones. Code is available at https://github.com/wangzy22/UniPre3D.
****
Charm: The Missing Piece in ViT Fine-Tuning for Image Aesthetic Assessment
Fatemeh Behrad, Tinne Tuytelaars, Johan Wagemans; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7815-7824
The capacity of Vision transformers (ViTs) to handle variable-sized inputs is often constrained by computational complexity and batch processing limitations. Consequently, ViTs are typically trained on small, fixed-size images obtained through downscaling or cropping. While reducing computational burden, these methods result in significant information loss, negatively affecting tasks like image aesthetic assessment. We introduce Charm, a novel tokenization approach that preserves Composition, High-resolution, Aspect Ratio, and Multi-scale information simultaneously. Charm prioritizes high-resolution details in specific regions while downscaling others, enabling shorter fixed-size input sequences for ViTs while incorporating essential information. Charm is designed to be compatible with pre-trained ViTs and their learned positional embeddings. By providing multiscale input and introducing variety to input tokens, Charm improves ViT performance and generalizability for image aesthetic assessment. We avoid cropping or changing

the aspect ratio to further preserve information. Extensive experiments demonstrate significant performance improvements on various image aesthetic and quality assessment datasets (up to 8.1%) using a lightweight ViT backbone. Code and pre-trained models are available at https://github.com/FBehrad/Charm.

****

F-LMM: Grounding Frozen Large Multimodal Models

Size Wu, Sheng Jin, Wenwei Zhang, Lumin Xu, Wentao Liu, Wei Li, Chen Change Loy; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24710-24721

Endowing Large Multimodal Models (LMMs) with visual grounding capability can significantly enhance AIs' understanding of the visual world and their interaction with humans. However, existing methods typically fine-tune the parameters of LMMs to learn additional segmentation tokens and overfit grounding and segmentation datasets. Such a design would inevitably cause a catastrophic diminution in the indispensable conversational capability of general AI assistants. In this paper, we comprehensively evaluate state-of-the-art grounding LMMs across a suite of multimodal question-answering benchmarks, observing drastic performance drops that indicate vanishing general knowledge comprehension and weakened instruction following ability. To address this issue, we present F-LMM---grounding frozen off-the-shelf LMMs in human-AI conversations---a straightforward yet effective design based on the fact that word-pixel correspondences conducive to visual grounding inherently exist in the attention mechanism of well-trained LMMs. Using only a few trainable CNN layers, we can translate word-pixel attention weights to mask logits, which a SAM-based mask refiner can further optimise. Our F-LMM neither learns special segmentation tokens nor utilises high-quality grounded instruction-tuning data, but achieves competitive performance on referring expression segmentation and panoptic narrative grounding benchmarks while completely preserving LMMs' original conversational ability. Additionally, with instruction-following ability preserved and grounding ability obtained, F-LMM can be directly applied to complex tasks like reasoning segmentation, grounded conversation generation and visual chain-of-thought reasoning. Our code can be found at https://github.com/wusize/F-LMM.

****

EntityErasure: Erasing Entity Cleanly via Amodal Entity Segmentation and Completion

Yixing Zhu, Qing Zhang, Yitong Wang, Yongwei Nie, Wei-Shi Zheng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28274-28283

This paper presents EntityErasure, a novel diffusion-based inpainting method that can effectively erase entities without inducing unwanted sundries. To this end, we propose to address this problem by dividing it into amodal entity segmentation and completion, such that the region to inpaint takes only entities in the non-inpainting area as reference, avoiding the possibility to generate unpredictable sundries. Moreover, we develop two entity segmentation based metrics for quantitatively assessing the performance of object erasure, which are shown be more effective than existing metrics. Experimental results demonstrate that our approach outperforms other state-of-the-art object erasure methods. Our code and data are available at https://zyxunh.github.io/EntityErasure-ProjectPage/.

****

Generative Video Propagation

Shaoteng Liu, Tianyu Wang, Jui-Hsien Wang, Qing Liu, Zhifei Zhang, Joon-Young Lee, Yijun Li, Bei Yu, Zhe Lin, Soo Ye Kim, Jiaya Jia; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17712-17722

Large-scale video generation models have the inherent ability to realistically model natural scenes. In this paper, we demonstrate that through a careful design of a generative video propagation framework, various video tasks can be addressed in a unified way by leveraging the generative power of such models. Specifically, our framework, GenProp, encodes the original video with a selective content encoder and propagates the changes made to the first frame using an image-to-video generation model. We propose a data generation scheme to cover multiple vide

o tasks based on instance-level video segmentation datasets. Our model is trained by incorporating a mask prediction decoder head and optimizing a region-aware loss to aid the encoder to preserve the original content while the generation model propagates the modified region. This novel design opens up new possibilities: In editing scenarios, GenProp allows substantial changes to an object's shape; for insertion, the inserted objects can exhibit independent motion; for removal, GenProp effectively removes effects like shadows and reflections from the whole video; for tracking, GenProp is capable of tracking objects and their associated effects together. Experiment results demonstrate the leading performance of our model in various video tasks, and we further provide in-depth analyses of the proposed framework.

****

From Multimodal LLMs to Generalist Embodied Agents: Methods and Lessons

Andrew Szot, Bogdan Mazoure, Omar Attia, Aleksei Timofeev, Harsh Agrawal, Devon Hjelm, Zhe Gan, Zsolt Kira, Alexander Toshev; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10644-10655

We examine the capability of Multimodal Large Language Models (MLLMs) to tackle diverse domains that extend beyond the traditional language and vision tasks these models are typically trained on. Specifically, our focus lies in areas such as Embodied AI, Games, UI Control, and Planning. To this end, we introduce a process of adapting an MLLM to a Generalist Embodied Agent (GEA). GEA is a single unified model capable of grounding itself across these varied domains through a multi-embodiment action tokenizer. GEA is trained with supervised learning on a large dataset of embodied experiences and with online RL in interactive simulators. We explore the data and algorithmic choices necessary to develop such a model. Our findings reveal the importance of training with cross-domain data and online RL for building generalist agents. The final GEA model achieves strong generalization performance to unseen tasks across diverse benchmarks compared to other generalist models and benchmark-specific approaches.

****

Mosaic3D: Foundation Dataset and Model for Open-Vocabulary 3D Segmentation

Junha Lee, Chunghyun Park, Jaesung Choe, Yu-Chiang Frank Wang, Jan Kautz, Minsu Cho, Chris Choy; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14089-14101

We tackle open-vocabulary 3D scene segmentation tasks by introducing a novel data generation pipeline and training framework. Our work targets three essential aspects required for an effective dataset: precise 3D region segmentation, comprehensive textual descriptions, and sufficient dataset scale. By leveraging state-of-the-art open-vocabulary image segmentation models and region-aware vision-language models (VLM), we develop an automatic pipeline capable of producing high-quality 3D mask-text pairs. Applying this pipeline to multiple 3D scene datasets, we create Mosaic3D-5.6M, a dataset of more than 30K annotated scenes with 5.6M mask-text pairs - significantly larger than existing datasets. Building on these data, we propose Mosaic3D, a 3D visiual foundation model (3D-VFM) combining a 3D encoder trained with contrastive learning and a lightweight mask decoder for open-vocabulary 3D semantic and instance segmentation. Our approach achieves state-of-the-art results on open-vocabulary 3D semantic and instance segmentation benchmarks including ScanNet200, Matterport3D, and ScanNet++, with ablation studies validating the effectiveness of our large-scale training data. https://nvlabs.github.io/Mosaic3D/

****

T-CIL: Temperature Scaling using Adversarial Perturbation for Calibration in Class-Incremental Learning

Seong-Hyeon Hwang, Minsu Kim, Steven Euijong Whang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15339-15348

We study model confidence calibration in class-incremental learning, where models learn from sequential tasks with different class sets. While existing works primarily focus on accuracy, maintaining calibrated confidence has been largely overlooked. Unfortunately, most post-hoc calibration techniques are not designed to work with the limited memories of old-task data typical in class-incremental l

earning, as retaining a sufficient validation set would be impractical. Thus, we propose T-CIL, a novel temperature scaling approach for class-incremental learning without a validation set for old tasks, that leverages adversarially perturbed exemplars from memory. Directly using exemplars is inadequate for temperature optimization, since they are already used for training. The key idea of T-CIL is to perturb exemplars more strongly for old tasks than for the new task by adjusting the perturbation direction based on feature distance, with the single magnitude determined using the new-task validation set. This strategy makes the perturbation magnitude computed from the new task also applicable to old tasks, leveraging the tendency that the accuracy of old tasks is lower than that of the new task. We empirically show that T-CIL significantly outperforms various baselines in terms of calibration on real datasets and can be integrated with existing class-incremental learning techniques with minimal impact on accuracy.

****

LoRA Subtraction for Drift-Resistant Space in Exemplar-Free Continual Learning

Xuan Liu, Xiaobin Chang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15308-15318

In continual learning (CL), catastrophic forgetting often arises due to feature drift. This challenge is particularly prominent in the exemplar-free continual learning (EFCL) setting, where samples from previous tasks cannot be retained, making it difficult to preserve prior knowledge. To address this issue, some EFCL methods aim to identify feature spaces that minimize the impact on previous tasks while accommodating new ones. However, they rely on static features or outdated statistics stored from old tasks, which prevents them from capturing the dynamic evolution of the feature space in CL, leading to performance degradation over time. In this paper, we introduce the Drift-Resistant Space (DRS), which effectively handles feature drifts without requiring explicit feature modeling or the storage of previous tasks. A novel parameter-efficient fine-tuning approach called Low-Rank Adaptation Subtraction (LoRA^-) is proposed to develop the DRS. This method subtracts the LoRA weights of old tasks from the initial pre-trained weight before processing new task data to establish the DRS for model training. Therefore, LoRA^- enhances stability, improves efficiency, and simplifies implementation. Furthermore, stabilizing feature drifts allows for better plasticity by learning with a triplet loss. Our method consistently achieves state-of-the-art results, especially for long task sequences, across multiple datasets.

****

Joint Out-of-Distribution Filtering and Data Discovery Active Learning

Sebastian Schmidt, Leonard Schenk, Leo Schwinn, Stephan Günnemann; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25677-25687

As the data demand for deep learning models increases, active learning (AL) becomes essential to strategically select samples for labeling, which maximizes data efficiency and reduces training costs. Real-world scenarios necessitate the consideration of incomplete data knowledge within AL. Prior works address handling out-of-distribution (OOD) data, while another research direction has focused on category discovery. However, a combined analysis of real-world considerations combining AL with out-of-distribution data and category discovery remains unexplored. To address this gap, we propose Joint Out-of-distribution filtering and data Discovery Active learning (Joda), to uniquely address both challenges simultaneously by filtering out OOD data before selecting candidates for labeling. In contrast to previous methods, we deeply entangle the training procedure with filter and selection to construct a common feature space that aligns known and novel categories while separating OOD samples. Unlike previous works, Joda is highly efficient and completely omits auxiliary models and training access to the unlabeled pool for filtering or selection. In extensive experiments on 18 configurations and 3 metrics, Joda consistently achieves the highest accuracy with the best class discovery to OOD filtering balance compared to state-of-the-art competitor approaches.

****

AniMer: Animal Pose and Shape Estimation Using Family Aware Transformer

Jin Lyu, Tianyi Zhu, Yi Gu, Li Lin, Pujin Cheng, Yebin Liu, Xiaoying Tang, Liang An; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17486-17496

Quantitative analysis of animal behavior and biomechanics requires accurate animal pose and shape estimation across species, and is important for animal welfare and biological research. However, the small network capacity of previous methods and limited multi-species dataset leave this problem underexplored. To this end, this paper presents AniMer to estimate animal pose and shape using family-aware Transformer, enhancing the reconstruction accuracy of diverse quadrupedal families. A key insight of AniMer is its integration of a high-capacity Transformer-based backbone and an animal family supervised contrastive learning scheme, unifying the discriminative understanding of various quadrupedal shapes within a single framework. For effective training, we aggregate most available open-sourced quadrupedal datasets, either with 3D or 2D labels. To improve the diversity of 3D labeled data, we introduce CtrlAni3D, a novel large-scale synthetic dataset created through a new diffusion-based conditional image generation pipeline. CtrlAni3D consists of about 10k images with pixel-aligned SMAL labels. In total, we obtain 41.3k annotated images for training and validation. Consequently, the combination of a family aware Transformer network and an expansive dataset enables AniMer to outperform existing methods not only on 3D datasets like Animal3D and CtrlAni3D, but also on out-of-distribution Animal Kingdom dataset. Ablation studies further demonstrate the effectiveness of our network design and CtrlAni3D in enhancing the performance of AniMer for in-the-wild applications. Project page: https://luoxuestar.github.io/AniMer_project_page/.

****

Co-op: Correspondence-based Novel Object Pose Estimation

Sungphill Moon, Hyeontae Son, Dongcheol Hur, Sangwook Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11622-11632

We propose Co-op, a novel method for accurately and robustly estimating the 6DoF pose of objects unseen during training from a single RGB image. Our method requires only the CAD model of the target object and can precisely estimate its pose without any additional fine-tuning. While existing model-based methods suffer from inefficiency due to using a large number of templates, our method enables fast and accurate estimation with a small number of templates. This improvement is achieved by finding semi-dense correspondences between the input image and the pre-rendered templates. Our method achieves strong generalization performance by leveraging a hybrid representation that combines patch-level classification and offset regression. Additionally, our pose refinement model estimates probabilistic flow between the input image and the rendered image, refining the initial estimate to an accurate pose using a differentiable PnP layer. We demonstrate that our method not only estimates object poses rapidly but also outperforms existing methods by a large margin on the seven core datasets of the BOP Challenge, achieving state-of-the-art accuracy.

****

Finding Local Diffusion Schrodinger Bridge using Kolmogorov-Arnold Network

Xingyu Qiu, Mengying Yang, Xinghua Ma, Fanding Li, Dong Liang, Gongning Luo, Wei Wang, Kuanquan Wang, Shuo Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23227-23236

In image generation, Schrodinger Bridge (SB)-based methods theoretically enhance the efficiency and quality compared to the diffusion models by finding the least costly path between two distributions. However, they are computationally expensive and time-consuming when applied to complex image data. The reason is that they focus on fitting globally optimal paths in high-dimensional spaces, directly generating images as next step on the path using complex networks through self-supervised training, which typically results in a gap with the global optimum. Meanwhile, most diffusion models are in the same path subspace generated by weights f_A(t) and f_B(t), as they follow the paradigm (x_t = f_A(t)x_ Img + f_B(t)epsilon). To address the limitations of SB-based methods, this paper proposes for the first time to find local Diffusion Schrodinger Bridges (LDSB) in the diffusion path subspace, which strengthens the connection between the SB problem and d

iffusion models. Specifically, our method optimizes the diffusion paths using Ko lmogorov-Arnold Network (KAN), which has the advantage of resistance to forgetti ng and continuous output. The experiment shows that our LDSB significantly impro ves the quality and efficiency of image generation using the same pre-trained de noising network and the KAN for optimising is only less than 0.1MB. The FID metr ic is reduced by more than 15%, especially with a reduction of 48.50% when NFE o f DDIM is 5 for the CelebA dataset. Code is available at https://github.com/Perc eptionComputingLab/LDSB.
****

CorrBEV: Multi-View 3D Object Detection by Correlation Learning with Multi-modal Prototypes
Ziteng Xue, Mingzhe Guo, Heng Fan, Shihui Zhang, Zhipeng Zhang; Proceedings of t he Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27413-27 423
Camera-only multi-view 3D object detection in autonomous driving has witnessed e ncouraging developments in recent years, largely attributed to the revolution of fundamental architectures in modeling bird's eye view (BEV). Despite the growin g overall average performance, we contend that the exploration of more specific and challenging corner cases hasn't received adequate attention. In this work, w e delve into a specific yet critical issue for safe autonomous driving: occlusio n. To alleviate this challenge, we draw inspiration from the human amodal percep tion system, which is proven to have the capacity for mentally reconstructing th e complete semantic concept of occluded objects with prior knowledge. More speci fically, we introduce auxiliary visual and language prototypes, akin to human pr ior knowledge, to enhance the diminished object features caused by occlusion. In spired by Siamese object tracking, we fuse the information from these prototypes with the baseline model through an efficient depth-wise correlation, thereby en hancing the quality of object-related features and guiding the learning of 3D ob ject queries, especially for partially occluded ones. Furthermore, we propose th e random pixel drop to mimic occlusion and the multi-modal contrastive loss to a lign visual features of different occlusion levels to a unified space during tra ining. Our inspiration originates from addressing occlusion, however, we are sur prised to find that the proposed framework also enhances robustness in various c hallenging scenarios that diminish object representation, such as inclement weat her conditions. By applying our model to different baselines, i.e., BEVFormer an d SparseBEV, we demonstrate consistent improvements.
****

Completion as Enhancement: A Degradation-Aware Selective Image Guided Network fo r Depth Completion
Zhiqiang Yan, Zhengxue Wang, Kun Wang, Jun Li, Jian Yang; Proceedings of the Com puter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26943-26953
In this paper, we introduce the Selective Image Guided Network (SigNet), a nove l degradation-aware framework that transforms depth completion into depth enhanc ement for the first time. Moving beyond direct completion using convolutional ne ural networks (CNNs), SigNet initially densifies sparse depth data through non-C NN densification tools to obtain coarse yet dense depth. This approach eliminate s the mismatch and ambiguity caused by direct convolution over irregularly sampl ed sparse data. Subsequently, SigNet redefines completion as enhancement, establ ishing a self-supervised degradation bridge between the coarse depth and the tar geted dense depth for effective RGB-D fusion. To achieve this, SigNet leverages the implicit degradation to adaptively select high-frequency components (e.g., e dges) of RGB data to compensate for the coarse depth. This degradation is furthe r integrated into a multi-modal conditional Mamba, dynamically generating the st ate parameters to enable efficient global high-frequency information interaction . We conduct extensive experiments on the NYUv2, DIML, SUN RGBD, and TOFDC datas ets, demonstrating the state-of-the-art (SOTA) performance of SigNet.
****

CATANet: Efficient Content-Aware Token Aggregation for Lightweight Image Super-R esolution
Xin Liu, Jie Liu, Jie Tang, Gangshan Wu; Proceedings of the Computer Vision and

Pattern Recognition Conference (CVPR), 2025, pp. 17902-17912
Transformer-based methods have demonstrated impressive performance in low-level visual tasks such as Image Super-Resolution (SR). However, its computational complexity grows quadratically with the spatial resolution. A series of works attempt to alleviate this problem by dividing Low-Resolution images into local windows, axial stripes, or dilated windows. SR typically leverages the redundancy of images for reconstruction, and this redundancy appears not only in local regions but also in long-range regions. However, these methods limit attention computation to content-agnostic local regions, limiting directly the ability of attention to capture long-range dependency. To address these issues, we propose a lightweight Content-Aware Token Aggregation Network (CATANet). Specifically, we propose an efficient Content-Aware Token Aggregation module for aggregate long-range content-similar tokens, which shares token centers across all image tokens and updates them only during the training phase. Then we utilize intra-group self-attention to enable long-range information interaction. Moreover, we design an inter-group cross-attention to further enhance global information interaction. The experimental results show that, compared with the state-of-the-art cluster-based method SPIN, our method achieves superior performance, with a maximum PSNR improvement of 0.33dB and nearly double the inference speed.
****

SeeGround: See and Ground for Zero-Shot Open-Vocabulary 3D Visual Grounding
Rong Li, Shijie Li, Lingdong Kong, Xulei Yang, Junwei Liang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3707-3717
3D Visual Grounding (3DVG) aims to locate objects in 3D scenes based on textual descriptions, essential for applications like augmented reality and robotics. Traditional 3DVG approaches rely on annotated 3D datasets and predefined object categories, limiting scalability and adaptability. To overcome these limitations, we introduce SeeGround, a zero-shot 3DVG framework leveraging 2D Vision-Language Models (VLMs) trained on large-scale 2D data. SeeGround represents 3D scenes as a hybrid of query-aligned rendered images and spatially enriched text descriptions, bridging the gap between 3D data and 2D-VLMs input formats. We propose two modules: the Perspective Adaptation Module, which dynamically selects viewpoints for query-relevant image rendering, and the Fusion Alignment Module, which integrates 2D images with 3D spatial descriptions to enhance object localization. Extensive experiments on ScanRefer and Nr3D demonstrate that our approach outperforms existing zero-shot methods by large margins. Notably, we exceed weakly supervised methods and rival some fully supervised ones, outperforming previous SOTA by 7.7% on ScanRefer and 7.1% on Nr3D, showcasing its effectiveness in complex 3DVG task. Project website (with demo and code): https://seeground.github.io/.
****

RayFlow: Instance-Aware Diffusion Acceleration via Adaptive Flow Trajectories
Huiyang Shao, Xin Xia, Yuhong Yang, Yuxi Ren, Xing Wang, Xuefeng Xiao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18113-18123
Diffusion models have achieved remarkable success across various domains. However, their slow generation speed remains a critical challenge. Existing acceleration methods, while aiming to reduce steps, often compromise sample quality, controllability, or introduce training complexities. Therefore, we propose \underline RayFlow , a novel diffusion framework that addresses these limitations. Unlike previous methods, RayFlow guides each sample along a unique path towards an instance-specific target distribution. This method minimizes sampling steps while preserving generation diversity and stability. Furthermore, we introduce Time Sampler, an importance sampling technique to enhance training efficiency by focusing on crucial timesteps. Extensive experiments demonstrate RayFlow's superiority in generating high-quality images with improved speed, control, and training efficiency compared to existing acceleration techniques.
****

Linear Attention Modeling for Learned Image Compression
Donghui Feng, Zhengxue Cheng, Shen Wang, Ronghua Wu, Hongwei Hu, Guo Lu, Li Song ; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR),

2025, pp. 7623-7632
Recent years, learned image compression has made tremendous progress to achieve impressive coding efficiency. Its coding gain mainly comes from non-linear neural network-based transform and learnable entropy modeling. However, most studies focus on a strong backbone, and few studies consider a low complexity design. In this paper, we propose LALIC, a linear attention modeling for learned image compression. Specially, we propose to use Bi-RWKV blocks, by utilizing the Spatial Mix and Channel Mix modules to achieve more compact feature extraction, and apply the Conv based Omni-Shift module to adapt to two-dimensional latent representation. Furthermore, we propose a RWKV-based Spatial-Channel ConTeXt model (RWKV-SCCTX), that leverages the Bi-RWKV to modeling the correlation between neighboring features effectively. To our knowledge, our work is the first work to utilize efficient Bi-RWKV models with linear attention for learned image compression. Experimental results demonstrate that our method achieves competitive RD performances by outperforming VTM-9.1 by -15.26%, -15.41%, -17.63% in BD-rate on Kodak, CLIC and Tecnick datasets. The code is available at https://github.com/sjtu-medialab/RwkvCompress .
****

Around the World in 80 Timesteps: A Generative Approach to Global Visual Geolocation
Nicolas Dufour, Vicky Kalogeiton, David Picard, Loic Landrieu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23016-23026
Global visual geolocation predicts where an image was captured on Earth. Since images vary in how precisely they can be localized, this task inherently involves a significant degree of ambiguity. However, existing approaches are deterministic and overlook this aspect. In this paper, we aim to close the gap between traditional geolocalization and modern generative methods. We propose the first generative geolocation approach based on diffusion and Riemannian flow matching, where the denoising process operates directly on the Earth's surface. Our model achieves state-of-the-art performance on three visual geolocation benchmarks: OpenStreetView-5M, YFCC-100M, and iNat21. In addition, we introduce the task of probabilistic visual geolocation, where the model predicts a probability distribution over all possible locations instead of a single point. We introduce new metrics and baselines for this task, demonstrating the advantages of our diffusion-based approach. Codes and models will be made available.
****

Asynchronous Collaborative Graph Representation for Frames and Events
Dianze Li, Jianing Li, Xu Liu, Xiaopeng Fan, Yonghong Tian; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1655-1666
Integrating frames and events has become a widely accepted solution for various tasks in challenging scenarios. However, most multimodal methods directly convert events into image-like formats synchronized with frames and process each stream through separate two-branch backbones, making it difficult to fully exploit the spatiotemporal events while limiting inference frequency to the frame rate. To address these problems, we propose a novel asynchronous collaborative graph representation, namely ACGR, which is the first trial to explore a unified graph framework for asynchronously processing frames and events with high performance and low latency. Technically, we first construct unimodal graphs for frames and events to preserve their spatiotemporal properties and sparsity. Then, an asynchronous collaborative alignment module is designed to align and fuse frames and events into a unified graph and the ACGR is generated through graph convolutional networks. Finally, we innovatively introduce domain adaptation to enable cross-modal interactions between frames and events by aligning their feature spaces. Experimental results show that our approach outperforms state-of-the-art methods in both object detection and depth estimation tasks, while significantly reducing computational latency and achieving real-time inference up to 200 Hz. Our code can be available at https://github.com/dianzl/ACGR.
****

Real-time High-fidelity Gaussian Human Avatars with Position-based Interpolation

of Spatially Distributed MLPs
Youyi Zhan, Tianjia Shao, Yin Yang, Kun Zhou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26297-26307
Many works have succeeded in reconstructing Gaussian human avatars from multi-view videos. However, they either struggle to capture pose-dependent appearance details with a single MLP, or rely on a computationally intensive neural network to reconstruct high-fidelity appearance but with rendering performance degraded to non-real-time. We propose a novel Gaussian human avatar representation that can reconstruct high-fidelity pose-dependence appearance with details and meanwhile can be rendered in real time. Our Gaussian avatar is empowered by spatially distributed MLPs which are explicitly located on different positions on human body. The parameters stored in each Gaussian are obtained by interpolating from the outputs of its nearby MLPs based on their distances. To avoid undesired smooth Gaussian property changing during interpolation, for each Gaussian we define a set of Gaussian offset basis, and a linear combination of basis represents the Gaussian property offsets relative to the neutral properties. Then we propose to let the MLPs output a set of coefficients corresponding to the basis. In this way, although Gaussian coefficients are derived from interpolation and change smoothly, the Gaussian offset basis is learned freely without constraints. The smoothly varying coefficients combined with freely learned basis can still produce distinctly different Gaussian property offsets, allowing the ability to learn high-frequency spatial signals. We further use control points to constrain the Gaussians distributed on a surface layer rather than allowing them to be irregularly distributed inside the body, to help the human avatar generalize better when animated under novel poses. Compared to the state-of-the-art method, our method achieves better appearance quality with finer details while the rendering speed is significantly faster under novel views and novel poses.
****
ReconDreamer: Crafting World Models for Driving Scene Reconstruction via Online Restoration
Chaojun Ni, Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Wenkang Qin, Guan Huang, Chen Liu, Yuyin Chen, Yida Wang, Xueyang Zhang, Yifei Zhan, Kun Zhan, Peng Jia, Xianpeng Lang, Xingang Wang, Wenjun Mei; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1559-1569
Closed-loop simulation is crucial for end-to-end autonomous driving. Existing sensor simulation methods (e.g., NeRF and 3DGS) reconstruct driving scenes based on conditions that closely mirror training data distributions. However, these methods struggle with rendering novel trajectories, such as lane changes. Recent work, DriveDreamer4D, has demonstrated that integrating world model knowledge alleviates these issues. Although the training-free integration approach is efficient, it still struggles to render larger maneuvers, such as multi-lane shifts.Therefore, we introduce ReconDreamer, which enhances driving scene reconstruction through incremental integration of world model knowledge. Specifically, based on the world model, DriveRestorer is proposed to mitigate ghosting artifacts via online restoration. Additionally, we propose the progressive data update strategy to ensure high-quality rendering for larger maneuvers. Notably, ReconDreamer is the first method to effectively render in large maneuvers (e.g., across multiple lanes, spanning up to 6 meters). Additionally, experimental results demonstrate that ReconDreamer outperforms Street Gaussians in the NTA-IoU, NTL-IoU, and FID, with a relative improvement by 24.87%, 6.72%, and 29.97%. Furthermore, ReconDreamer surpasses DriveDreamer4D with PVG during large maneuver rendering, as verified by a relative improvement of 195.87% in the NTA-IoU metric and a comprehensive user study.
****
RoboSense: Large-scale Dataset and Benchmark for Egocentric Robot Perception and Navigation in Crowded and Unstructured Environments
Haisheng Su, Feixiang Song, Cong Ma, Wei Wu, Junchi Yan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27446-27455
Reliable embodied perception from an egocentric perspective is challenging yet essential for autonomous navigation technology of intelligent mobile agents. With

the growing demand of social robotics, near-field scene understanding becomes a n important research topic in the areas of egocentric perceptual tasks related t o navigation in both crowded and unstructured environments. Due to the complexit y of environmental conditions and difficulty of surrounding obstacles owing to t runcation and occlusion, the perception capability under this circumstance is st ill inferior. To further enhance the intelligence of mobile robots, in this pape r, we setup an egocentric multi-sensor data collection platform based on 3 main types of sensors (Camera, LiDAR and Fisheye), which supports flexible sensor con figurations to enable dynamic sight of view from ego-perspective, capturing eith er near or farther areas. Meanwhile, a large-scale multimodal dataset is constru cted, named RoboSense, to facilitate egocentric robot perception. Specifically, RoboSense contains more than 133K synchronized data with 1.4M 3D bounding box an d IDs annotated in the full 360^ \circ view, forming 216K trajectories across 7 .6K temporal sequences. It has 270xand 18xas many annotations of surrounding obs tacles within near ranges as the previous datasets collected for autonomous driv ing scenarios such as KITTI and nuScenes. Moreover, we define a novel matching c riterion for near-field 3D perception and prediction metrics. Based on RoboSense , we formulate 6 popular tasks to facilitate the future research development, wh ere the detailed analysis as well as benchmarks are also provided accordingly. D ata desensitization measures have been conducted for privacy protection.

****

Self-Supervised Large Scale Point Cloud Completion for Archaeological Site Resto ration

Aocheng Li, James R. Zimmer-Dauphinee, Rajesh Kalyanam, Ian Lindsay, Parker VanV alkenburgh, Steven Wernke, Daniel Aliaga; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11759-11768

Point cloud completion helps restore partial incomplete point clouds suffering o cclusions. Current self-supervised methods fail to give high fidelity completion for large objects with missing surfaces and unbalanced distribution of availabl e points. In this paper, we present a novel method for restoring large-scale poi nt clouds with limited and imbalanced ground-truth. Using rough boundary annotat ions for a region of interest, we project the original point clouds into a multi ple-center-of-projection (MCOP) image, where fragments are projected to images o f 5 channels (RGB, depth, and rotation). Completion of the original point cloud is reduced to inpainting the missing pixels in the MCOP images. Due to lack of c omplete structures and an unbalanced distribution of existing parts, we develop a self-supervised scheme which learns to infill the MCOP image with points resem bling existing "complete" patches. Special losses are applied to further enhance the regularity and consistency of completed MCOP images, which is mapped back t o 3D to form final restoration. Extensive experiments demonstrate the superiorit y of our method in completing 600+ incomplete and unbalanced archaeological stru ctures in Peru.

****

Chain of Attack: On the Robustness of Vision-Language Models Against Transfer-Ba sed Adversarial Attacks

Peng Xie, Yequan Bie, Jianda Mao, Yangqiu Song, Yang Wang, Hao Chen, Kani Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 20 25, pp. 14679-14689

Pre-trained vision-language models (VLMs) have showcased remarkable performance in image and natural language understanding, such as image captioning and respon se generation. As the practical applications of VLMs become increasingly widespr ead, their potential safety and robustness issues raise concerns that adversarie s may evade the system and cause these models to generate toxic content through malicious attacks. Therefore, evaluating the robustness of open-source VLMs agai nst adversarial attacks has garnered growing attention, with transfer-based atta cks as a representative black-box attacking strategy. However, most existing tra nsfer-based attacks neglect the importance of the semantic correlations between vision and text modalities, leading to sub-optimal adversarial example generatio n and attack performance. To address this issue, we present Chain of Attack (CoA), which iteratively enhances the generation of adversarial examples based on th

e multi-modal semantic update using a series of intermediate attacking steps, ac
hieving superior adversarial transferability and efficiency. A unified attack su
ccess rate computing method is further proposed for automatic evasion evaluation
. Extensive experiments conducted under the most realistic and high-stakes scena
rio, demonstrate that our attacking strategy is able to effectively mislead mode
ls to generate targeted responses using only black-box attacks without any knowl
edge of the victim models. The comprehensive robustness evaluation in our paper
provides insight into the vulnerabilities of VLMs and offers a reference for the
 safety considerations of future model developments.
****

Rate-In: Information-Driven Adaptive Dropout Rates for Improved Inference-Time U
ncertainty Estimation

Tal Zeevi, Ravid Shwartz-Ziv, Yann LeCun, Lawrence H. Staib, John A. Onofrey; Pr
oceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025
, pp. 20757-20766

Accurate uncertainty estimation is crucial for deploying neural networks in risk
-sensitive applications such as medical diagnosis. Monte Carlo Dropout is a wide
ly used technique for approximating predictive uncertainty by performing stochas
tic forward passes with dropout during inference. However, using static dropout
rates across all layers and inputs can lead to suboptimal uncertainty estimates,
 as it fails to adapt to the varying characteristics of individual inputs and ne
twork layers. Existing approaches optimize dropout rates during training using l
abeled data, resulting in fixed inference-time parameters that cannot adjust to
new data distributions, compromising uncertainty estimates in Monte Carlo simula
tions. In this paper, we propose Rate-In, an algorithm that dynamically adjusts
dropout rates during inference by quantifying the information loss induced by dr
opout in each layer's feature maps. By treating dropout as controlled noise inje
ction and leveraging information-theoretic principles, Rate-In adapts dropout ra
tes per layer and per input instance without requiring ground truth labels. By q
uantifying the functional information loss in feature maps, we adaptively tune d
ropout rates to maintain perceptual quality across diverse medical imaging tasks
 and architectural configurations. Our extensive empirical study on synthetic da
ta and real-world medical imaging tasks demonstrates that Rate-In improves calib
ration and sharpens uncertainty estimates compared to fixed or heuristic dropout
 rates without compromising predictive performance. Rate-In offers a practical,
unsupervised, inference-time approach to optimizing dropout for more reliable pr
edictive uncertainty estimation in critical applications.
****

Thin-Shell-SfT: Fine-Grained Monocular Non-rigid 3D Surface Tracking with Neural
 Deformation Fields

Navami Kairanda, Marc Habermann, Shanthika Naik, Christian Theobalt, Vladislav G
olyanik; Proceedings of the Computer Vision and Pattern Recognition Conference (
CVPR), 2025, pp. 11373-11383

3D reconstruction of highly deformable surfaces (e.g. cloths) from monocular RGB
 videos is a challenging problem, and no solution provides a consistent and accu
rate recovery of fine-grained surface details. To account for the ill-posed natu
re of the setting, existing methods use deformation models with statistical, neu
ral, or physical priors. They also predominantly rely on nonadaptive discrete su
rface representations (e.g. polygonal meshes), perform frame-by-frame optimisati
on leading to error propagation, and suffer from poor gradients of the mesh-base
d differentiable renderers. Consequently, fine surface details such as cloth wri
nkles are often not recovered with the desired accuracy. In response to these li
mitations, we propose Thin-Shell-SfT, a new method for non-rigid 3D tracking tha
t represents a surface as an implicit and continuous spatiotemporal neural field
. We incorporate continuous thin shell physics prior based on the Kirchhoff-Love
 model for spatial regularisation, which starkly contrasts the discretised alter
natives of earlier works. Lastly, we leverage 3D Gaussian splatting to different
iably render the surface into image space and optimise the deformations based on
 analysis-by-synthesis principles. Our Thin-Shell-SfT outperforms prior works qu
alitatively and quantitatively thanks to our continuous surface formulation in c

onjunction with a specially tailored simulation prior and surface-induced 3D Gau
ssians. See our project page at https://4dqv.mpi-inf.mpg.de/ThinShellSfT.
****

DeCLIP: Decoupled Learning for Open-Vocabulary Dense Perception
Junjie Wang, Bin Chen, Yulin Li, Bin Kang, Yichi Chen, Zhuotao Tian; Proceedings
 of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 148
24-14834
Dense visual prediction tasks have been constrained by their reliance on predefi
ned categories, limiting their applicability in real-world scenarios where visua
l concepts are unbounded. While Vision-Language Models (VLMs) like CLIP have sho
wn promise in open-vocabulary tasks, their direct application to dense predictio
n often leads to suboptimal performance due to limitations in local feature repr
esentation. In this work, we present our observation that CLIP's image tokens st
ruggle to effectively aggregate information from spatially or semantically relat
ed regions, resulting in features that lack local discriminability and spatial c
onsistency. To address this issue, we propose DeCLIP, a novel framework that enh
ances CLIP by decoupling the self-attention module to obtain "content" and "cont
ext" features respectively. The "content" features are aligned with image crop r
epresentations to improve local discriminability, while "context" features learn
 to retain the spatial correlations under the guidance of vision foundation mode
ls, such as DINO. Extensive experiments demonstrate that DeCLIP significantly ou
tperforms existing methods across multiple open-vocabulary dense prediction task
s, including object detection and semantic segmentation. Code is available at ht
tps://github.com/xiaomoguhz/DeCLIP.
****

SocialGesture: Delving into Multi-person Gesture Understanding
Xu Cao, Pranav Virupaksha, Wenqi Jia, Bolin Lai, Fiona Ryan, Sangmin Lee, James
M. Rehg; Proceedings of the Computer Vision and Pattern Recognition Conference (
CVPR), 2025, pp. 19509-19519
Previous research in human gesture recognition has largely overlooked multi-pers
on interactions, which are crucial for understanding the social context of natur
ally occurring gestures. This limitation in existing datasets presents a signifi
cant challenge in aligning human gestures with other modalities like language an
d speech. To address this issue, we introduce SocialGesture, the first large-sca
le dataset specifically designed for multi-person gesture analysis. SocialGestur
e features a diverse range of natural scenarios and supports multiple gesture an
alysis tasks, including video-based recognition and temporal localization, provi
ding a valuable resource for advancing the study of gesture during complex socia
l interactions. Furthermore, we propose a novel visual question answering (VQA)
task to benchmark vision language models' (VLMs) performance on social gesture u
nderstanding. Our findings highlight several limitations of current gesture reco
gnition models, offering insights into future directions for improvement in this
 field. SocialGesture is available at huggingface.co/datasets/IrohXu/SocialGestu
re.
****

GenFusion: Closing the Loop between Reconstruction and Generation via Videos
Sibo Wu, Congrong Xu, Binbin Huang, Andreas Geiger, Anpei Chen; Proceedings of t
he Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6078-608
8
Recently, 3D reconstruction and generation have demonstrated impressive novel vi
ew synthesis results, achieving high fidelity and efficiency. However, a notabl
e conditioning gap can be observed between these two fields, e.g. , scalable 3D
scene reconstruction often requires densely captured views, whereas 3D generatio
n typically relies on a single or no input view, which significantly limits thei
r applications. We found that the source of this phenomenon lies in the misalign
ment between 3D constraints and generative priors. To address this problem, we p
ropose a reconstruction-driven video diffusion model that learns to condition vi
deo frames on artifact-prone RGB-D renderings. Moreover, we propose a cyclical
fusion pipeline that iteratively adds restoration frames from the generative mod
el to the training set, enabling progressive expansion and addressing the viewpo

int saturation limitations seen in previous reconstruction and generation pipelines. Our evaluation, including view synthesis from sparse view and masked input, validates the effectiveness of our approach.

****

The PanAf-FGBG Dataset: Understanding the Impact of Backgrounds in Wildlife Behaviour Recognition

Otto Brookes, Maksim Kukushkin, Majid Mirmehdi, Colleen Stephens, Paula Dieguez, Thurston C. Hicks, Sorrel Jones, Kevin Lee, Maureen S. McCarthy, Amelia Meier, Emmanuelle Normand, Erin G. Wessling, Roman M. Wittig, Kevin Langergraber, Klaus Zuberbühler, Lukas Boesch, Thomas Schmid, Mimi Arandjelovic, Hjalmar Kühl, Tilo Burghardt; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5433-5443

Computer vision analysis of camera trap video footage is essential for wildlife conservation, as captured behaviours offer some of the earliest indicators of changes in population health. Recently, several high-impact animal behaviour datasets and methods have been introduced to encourage their use; however, the role of behaviour-correlated background information and its significant effect on out-of-distribution generalisation remain unexplored. In response, we present the PanAf-FGBG dataset, featuring 21 hours of wild chimpanzee behaviours, recorded at 389 individual camera locations. Uniquely, it pairs every video with a chimpanzee (referred to as a foreground video) with a corresponding background video (with no chimpanzee) from the same camera location. We present two views of the dataset: one with overlapping camera locations and one with disjoint locations. This setup enables, for the first time, direct evaluation of in-distribution and out-of-distribution conditions, and for the impact of backgrounds on behaviour recognition models to be quantified. All clips have rich behavioural annotations and metadata, including unique camera IDs. Additionally, we establish several baselines and present a highly effective latent-space normalisation technique that boosts out-of-distribution performance by +5.42% mAP for convolutional and +3.75% mAP for transformer-based models. Finally, we provide an in-depth analysis on the role of backgrounds in out-of-distribution behaviour recognition, including the so far unexplored impact of background durations. (i.e., the count of background frames within foreground videos). The dataset is available at https://obrookes.github.io/panaf-fgbg.github.io

****

Multi-modal Topology-embedded Graph Learning for Spatially Resolved Genes Prediction from Pathology Images with Prior Gene Similarity Information

Hang Shi, Changxi Chi, Peng Wan, Daoqiang Zhang, Wei Shao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20810-20819

The rapid development of spatial transcriptomics (ST) allows researchers to measure the spatial-level gene expression in tissues. Although powerful, the cost for collecting the ST data is expensive, and thus several studies aim to predict gene expression in ST by utilizing their corresponding H/E stained pathology images. The existing ST based gene expression prediction models either adopt the pre-trained networks or rely on the handcrafted features to describe the pathology images, which still lack a systematic way to combine them together to define a spot-level representation that can reflect the topological profiles of different spots. On the other hand, all the ST based gene prediction models treat the prediction task for each gene independently, which overlook the fact that the exploration of potential interrelationships among them can help improve the prediction performance for individual genes.To address the above issues, we propose a multi-modal topology-embedded graph learning algorithm guided by prior Gene Ontology similarity information (i.e., M2TGLGO) to predict the spatial resolved genes from pathology image. Specifically, M2TGLGO co-learns the image representation of different spots from both deep and handcrafted features by considering the within-modal and inter-modal interactions. Next, to keep the topological structure among different spots, a spatial-oriented ranking module is also incorporated to preserve their neighborhood similarity information. Finally, we present a Gene Ontology knowledge guided graph neural network for simultaneously predicting multiple gene expressions by considering their functional associations. We evaluate

our method on three public available ST datasets, the experimental results show the effectiveness of our M2TGLGO in comparison with the existing studies.
****

Question-Aware Gaussian Experts for Audio-Visual Question Answering
Hongyeob Kim, Inyoung Jung, Dayoon Suh, Youjia Zhang, Sangmin Lee, Sungeun Hong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13681-13690
Audio-Visual Question Answering (AVQA) requires not only question-based multimodal reasoning but also precise temporal grounding to capture subtle dynamics for accurate prediction. However, existing methods mainly use question information implicitly, limiting focus on question-specific details. Furthermore, most studies rely on uniform frame sampling, which can miss key question-relevant frames. Although recent Top-K frame selection methods aim to address this, their discrete nature still overlooks fine-grained temporal details. This paper proposes QA-TIGER, a novel framework that explicitly incorporates question information and models continuous temporal dynamics. Our key idea is to use Gaussian-based modeling to adaptively focus on both consecutive and non-consecutive frames based on the question, while explicitly injecting question information and applying progressive refinement. We leverage a Mixture of Experts (MoE) to flexibly implement multiple Gaussian models, activating temporal experts specifically tailored to the question. Extensive experiments on multiple AVQA benchmarks show that QA-TIGER consistently achieves state-of-the-art performance.
****

Sonic: Shifting Focus to Global Audio Perception in Portrait Animation
Xiaozhong Ji, Xiaobin Hu, Zhihong Xu, Junwei Zhu, Chuming Lin, Qingdong He, Jiangning Zhang, Donghao Luo, Yi Chen, Qin Lin, Qinglin Lu, Chengjie Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 193-203
The study of talking face generation mainly explores the intricacies of synchronizing facial movements and crafting visually appealing, temporally-coherent animations. However, due to the limited exploration of global audio perception, current approaches predominantly employ auxiliary visual and spatial knowledge to stabilize the movements, which often results in the deterioration of the naturalness and temporal inconsistencies.Considering the essence of audio-driven animation, the audio signal serves as the ideal and unique priors to adjust facial expressions and lip movements, without resorting to interference of any visual signals. Based on this motivation, we propose a novel paradigm, dubbed as Sonic, to shift focus on the exploration of global audio perception. To effectively leverage global audio knowledge, we disentangle it into intra- and inter-clip audio perception and collaborate with both aspects to enhance overall perception. For the intra-clip audio perception, 1). Context-enhanced audio learning, in which long-range intra-clip temporal audio knowledge is extracted to provide facial expression and lip motion priors implicitly expressed as the tone and speed of speech. 2). Motion-decoupled Controller, in which the motion of the head and expression movement are disentangled and independently controlled by intra-audio clips. Most importantly, for inter-clip audio perception, as a bridge to connect the intra-clips to achieve the global perception, Time-aware position shift fusion, in which the global inter-clip audio information is considered and fused for long-audio inference via through consecutively time-aware shifted windows. Extensive experiments demonstrate that the novel audio-driven paradigm outperform existing SOTA methodologies in terms of video quality, temporally consistency, lip synchronization precision, and motion diversity.
****

Multitwine: Multi-Object Compositing with Text and Layout Control
Gemma Canet Tarrés, Zhe Lin, Zhifei Zhang, He Zhang, Andrew Gilbert, John Collomosse, Soo Ye Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8094-8104
We introduce the first generative model capable of simultaneous multi-object compositing, guided by both text and layout. Our model allows for the addition of multiple objects within a scene, capturing a range of interactions from simple p

ositional relations (e.g., next to, in front of) to complex actions requiring re posing (e.g., hugging, playing guitar). When an interaction implies additional p rops, like 'taking a selfie', our model autonomously generates these supporting objects. By jointly training for compositing and subject-driven generation, also known as customization, we achieve a more balanced integration of textual and v isual inputs for text-driven object compositing. As a result, we obtain a versat ile model with state-of-the-art performance in both tasks. We further present a data generation pipeline leveraging visual and language models to effortlessly s ynthesize multimodal, aligned training data.
****

DEFOM-Stereo: Depth Foundation Model Based Stereo Matching
Hualie Jiang, Zhiqiang Lou, Laiyan Ding, Rui Xu, Minglang Tan, Wenjie Jiang, Rui Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (C VPR), 2025, pp. 21857-21867
Stereo matching is a key technique for metric depth estimation in computer visio n and robotics. Real-world challenges like occlusion and non-texture hinder accu rate disparity estimation from binocular matching cues. Recently, monocular rela tive depth estimation has shown remarkable generalization using vision foundatio n models. Thus, to facilitate robust stereo matching with monocular depth cues, we incorporate a robust monocular relative depth model into the recurrent stereo -matching framework, building a new framework for depth foundation model-based s tereo-matching, DEFOM-Stereo. In the feature extraction stage, we construct the combined context and matching feature encoder by integrating features from conve ntional CNNs and DEFOM. In the update stage, we use the depth predicted by DEFOM to initialize the recurrent disparity and introduce a scale update module to re fine the disparity at the correct scale. DEFOM-Stereo is verified to have much s tronger zero-shot generalization compared with SOTA methods. Moreover, DEFOM-Ste reo achieves top performance on the KITTI 2012, KITTI 2015, Middlebury, and ETH3 D benchmarks, ranking 1^ st on many metrics. In the joint evaluation under the robust vision challenge, our model simultaneously outperforms previous models on the individual benchmarks, further demonstrating its outstanding capabilities.
****

Adaptive Rectangular Convolution for Remote Sensing Pansharpening
Xueyang Wang, Zhixin Zheng, Jiandong Shao, Yule Duan, Liang-Jian Deng; Proceedin gs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1 7872-17881
Recent advancements in convolutional neural network (CNN)-based techniques for r emote sensing pansharpening have markedly enhanced image quality. However, conve ntional convolutional modules in these methods have two critical drawbacks. Firs t, the sampling positions in convolution operations are confined to a fixed squa re window. Second, the number of sampling points is preset and remains unchanged . Given the diverse object sizes in remote sensing images, these rigid parameter s lead to suboptimal feature extraction. To overcome these limitations, we intro duce an innovative convolutional module, Adaptive Rectangular Convolution (ARCon v). ARConv adaptively learns both the height and width of the convolutional kern el and dynamically adjusts the number of sampling points based on the learned sc ale. This approach enables ARConv to effectively capture scale-specific features of various objects within an image, optimizing kernel sizes and sampling locati ons. Additionally, we propose ARNet, a network architecture in which ARConv is t he primary convolutional module. Extensive evaluations across multiple datasets reveal the superiority of our method in enhancing pansharpening performance over previous techniques. Ablation studies and visualization further confirm the eff icacy of ARConv. The source code can be available at https://github.com/WangXue yang-uestc/ARConv.
****

Video Depth without Video Models
Bingxin Ke, Dominik Narnhofer, Shengyu Huang, Lei Ke, Torben Peters, Katerina Fr agkiadaki, Anton Obukhov, Konrad Schindler; Proceedings of the Computer Vision a nd Pattern Recognition Conference (CVPR), 2025, pp. 7233-7243
Video depth estimation lifts monocular video clips to 3D by inferring dense dept

h at every frame. Recent advances in single-image depth estimation, brought abou t by the rise of large foundation models and the use of synthetic training data, have fueled a renewed interest in video depth. However, naively applying a sing le-image depth estimator to every frame of a video disregards temporal continuit y, which not only leads to flickering but may also break when camera motion caus es sudden changes in depth range. An obvious and principled solution would be to build on top of video foundation models, but these come with their own limitati ons; including expensive training and inference, imperfect 3D consistency, and s titching routines for the fixed-length (short) outputs. We take a step back and demonstrate how to turn a single-image latent diffusion model (LDM) into a state -of-the-art video depth estimator. Our model, which we call RollingDepth, has tw o main ingredients: (i) a multi-frame depth estimator that is derived from a sin gle-image LDM and maps very short video snippets (typically frame triplets) to d epth snippets. (ii) a robust, optimization-based registration algorithm that opt imally assembles depth snippets sampled at various different frame rates back in to a consistent video. RollingDepth is able to efficiently handle long videos wi th hundreds of frames and delivers more accurate depth videos than both dedicate d video depth estimators and high-performing single-frame models. Project page: https://rollingdepth.github.io.

****

PointLoRA: Low-Rank Adaptation with Token Selection for Point Cloud Learning

Song Wang, Xiaolu Liu, Lingdong Kong, Jianyun Xu, Chunyong Hu, Gongfan Fang, Wen tong Li, Jianke Zhu, Xinchao Wang; Proceedings of the Computer Vision and Patter n Recognition Conference (CVPR), 2025, pp. 6605-6615

Self-supervised representation learning for point cloud has demonstrated effecti veness in improving pre-trained model performance across diverse tasks. However, as pre-trained models grow in complexity, fully fine-tuning them for downstream applications demands substantial computational and storage resources. Parameter -efficient fine-tuning (PEFT) methods offer a promising solution to mitigate the se resource requirements, yet most current approaches rely on complex adapter an d prompt mechanisms that increase tunable parameters. In this paper, we propose PointLoRA, a simple yet effective method that combines low-rank adaptation (LoRA) with multi-scale token selection to efficiently fine-tune point cloud models. Our approach embeds LoRA layers within the most parameter-intensive components o f point cloud transformers, reducing the need for tunable parameters while enhan cing global feature capture. Additionally, multi-scale token selection extracts critical local information to serve as prompts for downstream fine-tuning, effec tively complementing the global context captured by LoRA. The experimental resul ts across various pre-trained models and three challenging public datasets demon strate that our approach achieves competitive performance with only 3.43% of the trainable parameters, making it highly effective for resource-constrained appli cations. Source code is available at: https://github.com/songw-zju/PointLoRA.

****

HumanRig: Learning Automatic Rigging for Humanoid Character in a Large Scale Dat aset

Zedong Chu, Feng Xiong, Meiduo Liu, Jinzhi Zhang, Mingqi Shao, Zhaoxu Sun, Di Wa ng, Mu Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 304-313

With the rapid evolution of 3D generation algorithms, the cost of producing 3D h umanoid character models has plummeted, yet the field is impeded by the lack of a comprehensive dataset for automatic rigging--a pivotal step in character anima tion. Addressing this gap, we present HumanRig, the first large-scale dataset sp ecifically designed for 3D humanoid character rigging, encompassing 11,434 metic ulously curated T-posed meshes adhered to a uniform skeleton topology. Capitaliz ing on this dataset, we introduce an innovative, data-driven automatic rigging f ramework, which overcomes the limitations of GNN-based methods in handling compl ex AI-generated meshes. Our approach integrates a Prior-Guided Skeleton Estimato r (PGSE) module, which uses 2D skeleton joints to provide a preliminary 3D skele ton, and a Mesh-Skeleton Mutual Attention Network (MSMAN) that fuses skeleton fe atures with 3D mesh features extracted by a U-shaped point transformer. This ena

bles a coarse-to-fine 3D skeleton joint regression and a robust skinning estimation, surpassing previous methods in quality and versatility. This work not only remedies the dataset deficiency in rigging research but also propels the animation industry towards more efficient and automated character rigging pipelines.
****

GaussHDR: High Dynamic Range Gaussian Splatting via Learning Unified 3D and 2D Local Tone Mapping
Jinfeng Liu, Lingtong Kong, Bo Li, Dan Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5991-6000
High dynamic range (HDR) novel view synthesis (NVS) aims to reconstruct HDR scenes by leveraging multi-view low dynamic range (LDR) images captured at different exposure levels. Current training paradigms with 3D tone mapping often result in unstable HDR reconstruction, while training with 2D tone mapping reduces the model's capacity to fit LDR images. Additionally, the global tone mapper used in existing methods can impede the learning of both HDR and LDR representations. To address these challenges, we present GaussHDR, which unifies 3D and 2D local tone mapping through 3D Gaussian splatting. Specifically, we design a residual local tone mapper for both 3D and 2D tone mapping that accepts an additional context feature as input. We then propose combining the dual LDR rendering results from both 3D and 2D local tone mapping at the loss level. Finally, recognizing that different scenes may exhibit varying balances between the dual results, we introduce uncertainty learning and use the uncertainties for adaptive modulation. Extensive experiments demonstrate that GaussHDR significantly outperforms state-of-the-art methods in both synthetic and real-world scenarios. The project page for this paper is available at https://liujf1226.github.io/GaussHDR.
****

UIBDiffusion: Universal Imperceptible Backdoor Attack for Diffusion Models
Yuning Han, Bingyin Zhao, Rui Chu, Feng Luo, Biplab Sikdar, Yingjie Lao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19186-19196
Recent studies show that diffusion models (DMs) are vulnerable to backdoor attacks. Existing backdoor attacks impose unconcealed triggers (e.g., a gray box and eyeglasses) that contain evident patterns, rendering remarkable attack effects yet easy detection upon human inspection and defensive algorithms. While it is possible to improve stealthiness by reducing the strength of the backdoor, doing so can significantly compromise its generality and effectiveness. In this paper, we propose UIBDiffusion, the universal imperceptible backdoor attack for diffusion models, which allows us to achieve superior attack and generation performance while evading state-of-the-art defenses. We propose a novel trigger generation approach based on universal adversarial perturbations (UAPs) and reveal that such perturbations, which are initially devised for fooling pre-trained discriminative models, can be adapted as potent imperceptible backdoor triggers for DMs. We evaluate UIBDiffusion on multiple types of DMs with different kinds of samplers across various datasets and targets. Experimental results demonstrate that UIBDiffusion brings three advantages: 1) Universality, the imperceptible trigger is universal (i.e., image and model agnostic) where a single trigger is effective to any images and all diffusion models with different samplers; 2) Utility, it achieves comparable generation quality (e.g., FID) and even better attack success rate (i.e., ASR) at low poison rates compared to the prior works; and 3) Undetectability, UIBDiffusion is plausible to human perception and can bypass Elijah and TERD, the SOTA defenses against backdoors for DMs. We will release our backdoor triggers and code.
****

DiskVPS: Vanishing Point Detector via Hough Transform in a Disk Region
Jianping Wu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27049-27058
DiskVPS is a novel and robust vanishing point (VP) detection scheme based on Hough Transform (HT) over an image-plane-mapped disk region. In the DiskVPS, the image plane is first mapped into a disk region, which is then partitioned into concentrical ring-shaped subspaces. Each subspace is further partitioned into appro

ximately equal-probability Hough cells. By using individual edges rather than ed ge pairs as voters, DiskVPS can achieve high accuracy with extreme efficiency. I nvolving no calibration parameters, DiskVPS is particularly suitable for VP dete ction in uncalibrated images and thus can be applied in image calibration. A com parative experimental study demonstrates that the basic DiskVPS model without pa rameter optimization achieved significantly better performance over the SOTA in detection accuracy and processing speed with real-world images. The study also s hows that DiskVPS is robust against parameter changes.
****

Seeing Far and Clearly: Mitigating Hallucinations in MLLMs with Attention Causal Decoding
Feilong Tang, Chengzhi Liu, Zhongxing Xu, Ming Hu, Zile Huang, Haochen Xue, Ziya ng Chen, Zelin Peng, Zhiwei Yang, Sijin Zhou, Wenxue Li, Yulong Li, Wenxuan Song , Shiyan Su, Wei Feng, Jionglong Su, Mingquan Lin, Yifan Peng, Xuelian Cheng, Im ran Razzak, Zongyuan Ge; Proceedings of the Computer Vision and Pattern Recognit ion Conference (CVPR), 2025, pp. 26147-26159
Recent advancements in multimodal large language models (MLLMs) have significant ly improved performance in visual question answering. However, they often suffer from hallucinations. In this work, hallucinations are categorized into two main types: initial hallucinations and snowball hallucinations. We argue that adequa te contextual information can be extracted directly from the token interaction p rocess. Inspired by causal inference in decoding strategy, we propose to leverag e causal masks to establish information propagation between multimodal tokens. T he hypothesis is that insufficient interaction between those tokens may lead the model to rely on outlier tokens, overlooking dense and rich contextual cues. Th erefore, we propose to intervene in the propagation process by tackling outlier tokens to enhance in-context inference. With this goal, we present FarSight, a v ersatile plug-and-play decoding strategy to reduce attention interference from o utlier tokens merely by optimizing the causal mask. The heart of our method is e ffective token propagation. We design an attention register structure within the upper triangular matrix of the causal mask, dynamically allocating attention ca pture attention diverted to outlier tokens. Moreover, a positional awareness enc oding method with a diminishing masking rate is proposed, allowing the model to attend to further preceding tokens, especially for video sequence tasks. With ex tensive experiments, FarSight demonstrates significant hallucination-mitigating performance across different MLLMs on both image and video benchmarks, proving i ts effectiveness.
****

Towards Autonomous Micromobility through Scalable Urban Simulation
Wayne Wu, Honglin He, Chaoyuan Zhang, Jack He, Seth Z. Zhao, Ran Gong, Quanyi Li , Bolei Zhou; Proceedings of the Computer Vision and Pattern Recognition Confere nce (CVPR), 2025, pp. 27553-27563
Micromobility, which utilizes lightweight devices moving in urban public spaces - such as delivery robots and electric wheelchairs - emerges as a promising alte rnative to vehicular mobility. Current micromobility depends mostly on human man ual operation (in-person or remote control), which raises safety and efficiency concerns when navigating busy urban environments full of obstacles and pedestria ns. Assisting humans with AI agents in maneuvering micromobility devices present s a viable solution for enhancing safety and efficiency. In this work, we presen t a scalable urban simulation solution to advance autonomous micromobility. Firs t, we build URBAN-SIM -- a high-performance robot learning platform for large-sc ale training of embodied agents in interactive urban scenes. URBAN-SIM contains three critical modules: Hierarchical Urban Generation pipeline, Interactive Dyna mics Generation strategy, and Asynchronous Scene Sampling scheme, to improve the diversity, realism, and efficiency of robot learning in simulation. Then, we pr opose URBAN-BENCH -- a suite of essential tasks and benchmarks to gauge various capabilities of the AI agents in achieving autonomous micromobility. URBAN-BENCH includes eight tasks based on three core skills of the agents: Urban Locomotion , Urban Navigation, and Urban Traverse. We evaluate four robots with heterogeneo us embodiments, such as the wheeled and legged robots, across these tasks. Exper

iments on diverse terrains and urban structures reveal each robot's unique strengths and limitations. This work will be open-sourced and under sustainable maintenance to foster future research in autonomous micromobility.
****

FisherTune: Fisher-Guided Robust Tuning of Vision Foundation Models for Domain Generalized Segmentation

Dong Zhao, Jinlong Li, Shuang Wang, Mengyao Wu, Qi Zang, Nicu Sebe, Zhun Zhong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15043-15054

Vision Foundation Models (VFMs) excel in generalization due to large-scale pretraining, but fine-tuning them for Domain Generalized Semantic Segmentation (DGSS) while maintaining this ability remains a challenge. Existing approaches either selectively fine-tune parameters or freeze the VFMs and update only the adapters, both of which may underutilize the VFMs' full potential in DGSS tasks. We observe that domain-sensitive parameters in VFMs, arising from task and distribution differences, can hinder generalization. To address this, we propose FisherTune, a robust fine-tuning method guided by the Domain-Related Fisher Information Matrix (DR-FIM). DR-FIM measures parameter sensitivity across tasks and domains, enabling selective updates that preserve generalization and enhance DGSS adaptability. To stabilize DR-FIM estimation, FisherTune incorporates variational inference, treating parameters as Gaussian-distributed variables and leveraging pretrained priors. Extensive experiments show that FisherTune achieves superior cross-domain segmentation while maintaining generalization, outperforming both selective-parameter and adapter-based methods.
****

Language-Assisted Debiasing and Smoothing for Foundation Model-Based Semi-Supervised Learning

Na Zheng, Xuemeng Song, Xue Dong, Aashish Nikhil Ghosh, Liqiang Nie, Roger Zimmermann; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25708-25717

Recent studies have focused on introducing pre-trained foundation models into semi-supervised learning (SSL) tasks. Nevertheless, these foundation models can exhibit biases toward different classes and tend to generate imbalanced pseudo-labels for SSL. Thus, efforts have been made to introduce the logit adjustment offset to reduce the inherent bias in foundation models for SSL tasks.Despite their success, existing foundation model-based SSL methods face challenges: 1) unreliability in the estimated logit adjustment offset, 2) overlooking the potential of linguistic knowledge in capturing model biases and 3) fail to fully exploit the unlabeled samples. To address these issues, we propose Language-Assisted Debiasing and Smoothing framework, namely LADaS, for foundation model-based SSL. It consists of two components: 1) Language-assisted Pseudo-Label Debiasing (LPLD) to reduce biases in foundation models, and 2) Language-aware Pseudo-Label Smoothing (LPLS) to fully exploit low-confidence samples to facilitate SSL training. In particular, LPLD introduces a reliability score to dynamically assess the reliability of the logit adjustment. Additionally, it incorporates a language-oriented preference to reduce model biases using linguistic knowledge derived from pre-trained language models. Finally, LPLS introduces language-aware soft labels and devises language-aware pseudo-label smoothing loss to guide the learning of unlabeled samples with low-quality pseudo-labels. Extensive experiments demonstrate the superiority of our proposed LADaS.
****

EdgeMovingNet: Edge-preserving Point Cloud Reconstruction via Joint Geometry Features

Xinran Yang, Donghao Ji, Yuanqi Li, Junyuan Xie, Jie Guo, Yanwen Guo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22150-22160

Point cloud reconstruction is a critical process in 3D representation and reverse engineering. When it comes to CAD models, edges are significant features that play a crucial role in characterizing the geometry of 3D shapes. However, few points are exactly sampled on edges during acquisition, resulting in apparent arti

facts for the reconstruction task. Upsampling point cloud is a direct technical route, but there is a main challenge that the upsampled points may not align with the model edge accurately. To overcome this, we develop an integrated framework to estimate edges by joint regression of three geometry features--point-to-edge direction, point-to-edge distance and point normal. Benefiting these features, we implement a novel refinement process to move and produce more points which lie accurately on edges of the model, allowing for high-quality edge-preserving reconstruction. Experiments and comparisons against previous methods demonstrate our method's effectiveness and superiority.
****

AdMiT: Adaptive Multi-Source Tuning in Dynamic Environments
Xiangyu Chang, Fahim Faisal Niloy, Sk Miraj Ahmed, Srikanth V. Krishnamurthy, Basak Guler, Ananthram Swami, Samet Oymak, Amit Roy-Chowdhury; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20569-20579
Incorporating transformer models into edge devices poses a significant challenge due to the computational demands of adapting these large models across diverse applications. Parameter-efficient tuning (PET) methods (e.g. LoRA, Adapter, Visual Prompt Tuning, etc.) allow for targeted adaptation by modifying only small parts of the transformer model. However, adapting to dynamic unlabeled target distributions at the test time remains complex. To address this, we introduce AdMiT: Adaptive Multi -Source Tuning in Dynamic Environments. AdMiT innovates by pre-training a set of PET modules, each optimized for different source distributions or tasks, and dynamically selecting and integrating a sparse subset of relevant modules when encountering a new, few-shot, unlabeled target distribution. This integration leverages Kernel Mean Embedding (KME)-based matching to align the target distribution with relevant source knowledge efficiently, without requiring additional routing networks or hyperparameter tuning. AdMiT achieves adaptation with a single inference step, making it particularly suitable for resource-constrained edge deployments. Furthermore, AdMiT preserves privacy by performing an adaptation locally on each edge device, without the need for data exchange. Our theoretical analysis establishes guarantees for AdMiT's generalization, while extensive benchmarks demonstrate that AdMiT consistently outperforms other PET methods across a range of tasks, achieving robust and efficient adaptation.
****

Unbiased Video Scene Graph Generation via Visual and Semantic Dual Debiasing
Yanjun Li, Zhaoyang Li, Honghui Chen, Lizhi Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19047-19056
Video Scene Graph Generation (VidSGG) aims to capture dynamic relationships among entities by sequentially analyzing video frames and integrating visual and semantic information. However, VidSGG is challenged by significant biases that skew predictions. To mitigate these biases, we propose a VIsual and Semantic Awareness (VISA) framework for unbiased VidSGG. VISA addresses visual bias through an innovative memory update mechanism that enhances object representations and concurrently reduces semantic bias by iteratively integrating object features with comprehensive semantic information derived from triplet relationships. This visual-semantics dual debiasing approach results in more unbiased representations of complex scene dynamics. Extensive experiments demonstrate the effectiveness of our method, where VISA outperforms existing unbiased VidSGG approaches by a substantial margin (e.g., +13.1% improvement in mR@20 and mR@50 for the SGCLS task under Semi Constraint).
****

Channel-wise Noise Scheduled Diffusion for Inverse Rendering in Indoor Scenes
JunYong Choi, Min-cheol Sagong, SeokYeong Lee, Seung-Won Jung, Ig-Jae Kim, Junghyun Cho; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5773-5782
We propose a diffusion-based inverse rendering framework that decomposes a single RGB image into geometry, material, and lighting. Inverse rendering is inherently ill-posed, making it difficult to predict a single accurate solution. To address this challenge, recent generative model-based methods aim to present a range of possible solutions. However, finding a single accurate solution and generati

ng diverse solutions can be conflicting. In this paper, we propose a channel-wis e noise scheduling approach that allows a single diffusion model architecture to achieve two conflicting objectives. The resulting two diffusion models, trained with different channel-wise noise schedules, can predict a single highly accura te solution and present multiple possible solutions. The experimental results de monstrate the superiority of our two models in terms of both diversity and accur acy, which translates to enhanced performance in downstream applications such as object insertion and material editing.
****

Targeted Forgetting of Image Subgroups in CLIP Models
Zeliang Zhang, Gaowen Liu, Charles Fleming, Ramana Rao Kompella, Chenliang Xu; P roceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 202 5, pp. 9870-9880
Foundation models (FMs) such as CLIP have demonstrated impressive zero-shot perf ormance across various tasks by leveraging large-scale, unsupervised pre-trainin g. However, they often inherit harmful or unwanted knowledge from noisy internet -sourced datasets, compromising their reliability in real-world applications. Ex isting model unlearning methods either rely on access to pre-trained datasets or focus on coarse-grained unlearning (e.g., entire classes), leaving a critical g ap for fine-grained unlearning. In this paper, we address the challenging scenar io of selectively forgetting specific portions of knowledge within a class--with out access to pre-trained data--while preserving the model's overall performance . We propose a novel three-stage approach that progressively unlearns targeted k nowledge while mitigating over-forgetting. It consists of (1) a forgetting stage to fine-tune the CLIP on samples to be forgotten, (2) a reminding stage to rest ore performance on retained samples, and (3) a restoring stage to recover zero-s hot capabilities using model souping. Additionally, we introduce knowledge disti llation to handle the distribution disparity between forgetting/retaining sample s and unseen pre-trained data. Extensive experiments on CIFAR-10, ImageNet-1K, a nd style datasets demonstrate that our approach effectively unlearns specific su bgroups while maintaining strong zero-shot performance on semantically similar s ubgroups and other categories, significantly outperforming baseline unlearning m ethods, which lose effectiveness under the CLIP unlearning setting.
****

Unleashing In-context Learning of Autoregressive Models for Few-shot Image Manip ulation
Bolin Lai, Felix Juefei-Xu, Miao Liu, Xiaoliang Dai, Nikhil Mehta, Chenguang Zhu , Zeyi Huang, James M. Rehg, Sangmin Lee, Ning Zhang, Tong Xiao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18346-1 8357
Text-guided image manipulation has experienced notable advancement in recent yea rs. In order to mitigate linguistic ambiguity, few-shot learning with visual exa mples has been applied for instructions that are underrepresented in the trainin g set, or difficult to describe purely in language. However, learning from visua l prompts requires strong reasoning capability, which diffusion models are strug gling with. To address this issue, we introduce a novel multi-modal autoregressi ve model, dubbed InstaManip, that can instantly learn a new image manipulation o peration from textual and visual guidance via in-context learning, and apply it to new query images. Specifically, we propose an innovative group self-attention mechanism to break down the in-context learning process into two separate stage s -- learning and applying, which simplifies the complex problem into two easier tasks. We also introduce a relation regularization method to further disentangl e image transformation features from irrelevant contents in exemplar images. Ext ensive experiments suggest that our method surpasses previous few-shot image man ipulation models by a notable margin (>=19% in human evaluation). We also find o ur model can be further boosted by increasing the number or diversity of exempla r images.
****

Harnessing Frozen Unimodal Encoders for Flexible Multimodal Alignment
Mayug Maniparambil, Raiymbek Akshulakov, Yasser Abdelaziz Dahou Djilali, Sanath

Narayan, Ankit Singh, Noel E. O'Connor; Proceedings of the Computer Vision and P
attern Recognition Conference (CVPR), 2025, pp. 29847-29857
Recent contrastive multimodal vision-language models like CLIP have demonstrated
 robust open-world semantic understanding, becoming the standard image backbones
 for vision-language applications. However, recent findings suggest high semanti
c similarity between well-trained unimodal encoders, which raises a key question
: Are semantically similar embedding spaces separated only by simple projection
transformations? To validate this, we propose a novel framework that aligns visi
on and language using frozen unimodal encoders. It involves selecting semantical
ly similar encoders in the latent space, curating a concept-rich dataset of imag
e-caption pairs, and training simple MLP projectors. We evaluated our approach o
n various tasks involving both strong unimodal vision (0-shot localization) and
language encoders (multi-lingual, long context) and show that simple Projectors
retain unimodal capabilities in joint embedding space. Furthermore, our best mod
el, utilizing DINOv2 and All-Roberta-Large text encoder, achieves 76(%) accuracy
 on ImageNet with a 20-fold reduction in data and 65-fold reduction in compute r
equirements compared to multimodal alignment where models are trained from scrat
ch. The proposed framework enhances the accessibility of multimodal model develo
pment while enabling flexible adaptation across diverse scenarios. Code and cura
ted datasets are available at github.com/mayug/freeze-align.
****

Feature Information Driven Position Gaussian Distribution Estimation for Tiny Ob
ject Detection
Jinghao Bian, Mingtao Feng, Weisheng Dong, Fangfang Wu, Jianqiao Luo, Yaonan Wan
g, Guangming Shi; Proceedings of the Computer Vision and Pattern Recognition Con
ference (CVPR), 2025, pp. 30376-30386
Tiny object detection remains challenging in spite of the success of generic det
ectors. The dramatic performance degradation of generic detectors on tiny object
s is mainly due to the the weak representations of extremely limited pixels. To
address this issue, we propose a plug-and-play architecture to enhance the extin
guished regions. We for the first time exploit the regions to be enhanced from t
he perspective of pixel-wise amount of information. Specifically, we model the e
ntire image pixels feature information by minimizing Information Entropy loss, g
enerating an information map to attentively highlight weak activated regions in
an unsupervised way. To effectively assist the above phase with more attention t
o tiny objects, we next introduce the Position Gaussian Distribution Map, explic
itly modeled using a Gaussian Mixture distribution, where each Gaussian componen
t's parameters depend on the position and size of object instance labels, servin
g as supervision for further feature enhancement. Taking the information map as
prior knowledge guidance, we construct a multi-scale position gaussian distribut
ion map prediction module, simultaneously modulating the information map and dis
tribution map to focus on tiny objects during training. Extensive experiments on
 three public tiny object datasets demonstrate the superiority of our method ove
r current state-of-the-art competitors.
****

Enhancing Diversity for Data-free Quantization
Kai Zhao, Zhihao Zhuang, Miao Zhang, Chenjuan Guo, Yang Shu, Bin Yang; Proceedin
gs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2
0969-20978
Model quantization is an effective way to compress deep neural networks and acce
lerate the inference time on edge devices. Existing quantization methods usually
 require original data for calibration during the compressing process, which may
 be inaccessible due to privacy issues. A common way is to generate calibration
data to mimic the origin data. However, the generators in these methods have the
 mode collapse problem, making them unable to synthesize diverse data. To solve
this problem, we leverage the information from the full-precision model and enha
nce both inter-class and intra-class diversity for generating better calibration
 data, by devising a multi-layer features mixer and normalization flow based att
ention. Besides, novel regulation losses are proposed to make the generator prod
uce diverse data with more patterns from the perspective of activated feature va

lues and for the quantized model to learn better clip ranges adaptive to our div
erse calibration data. Extensive experiments show that our method achieves state
-of-the-art quantization results for both Transformer and CNN architectures. In
addition, we visualize the generated data to verify that our strategies can effe
ctively handle the mode collapse issue. Our codes are available at https://anony
mous.4open.science/r/DFQ-84E6 and will be publicly available.
****

SeqAfford: Sequential 3D Affordance Reasoning via Multimodal Large Language Mode
l

Chunlin Yu, Hanqing Wang, Ye Shi, Haoyang Luo, Sibei Yang, Jingyi Yu, Jingya Wan
g; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR),
 2025, pp. 1691-1701

3D affordance segmentation aims to link human instructions to touchable regions
of 3D objects for embodied manipulations. Existing efforts typically adhere to s
ingle-object, single-affordance paradigms, where each affordance type or explici
t instruction strictly corresponds to a specific affordance region and are unabl
e to handle long-horizon tasks. Such a paradigm cannot actively reason about com
plex user intentions that often imply sequential affordances. In this paper, we
introduce the Sequential 3D Affordance Reasoning task, which extends the traditi
onal paradigm by reasoning from cumbersome user intentions and then decomposing
them into a series of segmentation maps. Toward this, we construct the first ins
truction-based affordance segmentation benchmark that includes reasoning over bo
th single and sequential affordances, comprising 180K instruction-point cloud pa
irs. Based on the benchmark, we propose our model, SeqAfford, to unlock the 3D m
ulti-modal large language model with additional affordance segmentation abilitie
s, which ensures reasoning with world knowledge and fine-grained affordance grou
nding in a cohesive framework. We further introduce a multi-granular language-po
int integration module to endow 3D dense prediction. Extensive experimental eval
uations show that our model excels over well-established methods and exhibits op
en-world generalization with sequential reasoning abilities.
****

DSV-LFS: Unifying LLM-Driven Semantic Cues with Visual Features for Robust Few-S
hot Segmentation

Amin Karimi, Charalambos Poullis; Proceedings of the Computer Vision and Pattern
 Recognition Conference (CVPR), 2025, pp. 4584-4594

Few-shot semantic segmentation (FSS) aims to enable models to segment novel/unse
en object classes using only a limited number of labeled examples. However, curr
ent FSS methods frequently struggle with generalization due to incomplete and bi
ased feature representations, especially when support images do not capture the
full appearance variability of the target class. To improve the FSS pipeline, we
 propose a novel framework that utilizes large language models (LLMs) to adapt g
eneral class semantic information to the query image. Furthermore, the framework
 employs dense pixel-wise matching to identify similarities between query and su
pport images, resulting in enhanced FSS performance. Inspired by reasoning-based
 segmentation frameworks, our method, named DSV-LFS, introduces an additional to
ken into the LLM vocabulary, allowing a multimodal LLM to generate a "semantic p
rompt" from class descriptions. In parallel, a dense matching module identifies
visual similarities between the query and support images, generating a "visual p
rompt". These prompts are then jointly employed to guide the prompt-based decode
r for accurate segmentation of the query image. Comprehensive experiments on the
 benchmark datasets Pascal-5i and COCO-20i demonstrate that our framework achiev
es state-of-the-art performance-by a significant margin-demonstrating superior g
eneralization to novel classes and robustness across diverse scenarios.
****

Revisiting Generative Replay for Class Incremental Object Detection

Shizhou Zhang, Xueqiang Lv, Yinghui Xing, Qirui Wu, Di Xu, Yanning Zhang; Procee
dings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp
. 20340-20349

Generative replay has gained significant attention in class-incremental learning
; however, its application to Class Incremental Object Detection (CIOD) remains

limited due to the challenges in generating complex images with precise spatial arrangements. In this study, motivated by the observation that the forgetting of prior knowledge is predominantly present in the classification sub-task as opposed to the localization sub-task, we revisit the generative replay method for class incremental object detection. Our method utilize a standard Stable Diffusion model to generate image-level replay data for all old and new tasks. Accordingly, the old detector and a stage-wise detector are conducted on the synthetic images respectively to determine the bounding box positions through pseudo-labeling. Furthermore, we propose to use a Similarity-based Cross Sampling mechanism to select more valuable confusing data between old and new tasks to more effectively mitigate catastrophic forgetting and reduce the false alarm rate for the new task. Finally, all synthetic and real data are integrated for current-stage detector training, where the images generated for previous tasks are highly beneficial in minimizing the forgetting of existing knowledge, while those synthesized for the new task can help bridge the domain gap between real and synthetic images. We conducted extensive experiments on PASCAL VOC 2007 and MS COCO benchmark datasets in multiple settings to showcase the efficacy of our proposed approach, which achieves state-of-the-art results. The code are available at https://github.com/qiangzai-lv/RGR-IOD.

****

SemAlign3D: Semantic Correspondence between RGB-Images through Aligning 3D Object-Class Representations

Krispin Wandel, Hesheng Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1138-1147

Semantic correspondence made tremendous progress through the recent advancements of large vision models (LVM). While these LVMs have been shown to reliably capture local semantics, the same can currently not be said for capturing global geometric relationships between semantic object regions. This problem leads to unreliable performance for semantic correspondence between images with extreme view variation. In this work, we aim to leverage monocular depth estimates to capture these geometric relationships for more robust and data-efficient semantic correspondence. First, we introduce a simple but effective method to build 3D object-class representations from monocular depth estimates and LVM features using a sparsely annotated image correspondence dataset. Second, we formulate an alignment energy that can be minimized using gradient descent to obtain an alignment between the 3D object-class representation and the object-class instance in the input RGB-image. Our method achieves state-of-the-art matching accuracy in multiple categories on the challenging SPair-71k dataset, increasing the PCK@0.1 score by more than 10 points on three categories and overall by 3.3 points from 85.6% to 88.9%. Additional resources and code are available at dub.sh/semalign3d.

****

Bridging Viewpoint Gaps: Geometric Reasoning Boosts Semantic Correspondence

Qiyang Qian, Hansheng Chen, Masayoshi Tomizuka, Kurt Keutzer, Qianqian Wang, Chenfeng Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11579-11589

Finding semantic correspondences between images is a challenging problem in computer vision, particularly under significant viewpoint changes. Previous methods rely on semantic features from pre-trained 2D models like Stable Diffusion and DINOv2, which often struggle to extract viewpoint-invariant features. To overcome this, we propose a novel approach that integrates geometric and semantic reasoning. Unlike prior methods relying on heuristic geometric enhancements, our framework fine-tunes DUSt3R on synthetic cross-instance data to reconstruct distinct objects in an aligned 3D space. By learning to deform these objects into similar shapes using semantic supervision, we enable efficient KNN-based geometric matching, followed by sparse semantic matching within local KNN candidates. While trained on synthetic data, our method generalizes effectively to real-world images, achieving up to 7.4-point improvements in zero-shot settings on the rigid-body subset of Spair-71K and up to 19.6-point gains under extreme viewpoint variations. Additionally, it accelerates runtime by up to 40 times, demonstrating both its robustness to viewpoint changes and its efficiency for practical applications

.
****
DPU: Dynamic Prototype Updating for Multimodal Out-of-Distribution Detection
Shawn Li, Huixian Gong, Hao Dong, Tiankai Yang, Zhengzhong Tu, Yue Zhao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10193-10202
Out-of-distribution (OOD) detection is crucial for ensuring the robustness of machine learning models by identifying samples that deviate from the training distribution. While traditional OOD detection has predominantly focused on single-modality inputs, such as images, recent advancements in multimodal models have shown the potential of utilizing multiple modalities (e.g., video, optical flow, audio) to improve detection performance. However, existing approaches often neglect intra-class variability within in-distribution (ID) data, assuming that samples of the same class are perfectly cohesive and consistent. This assumption can lead to performance degradation, especially when prediction discrepancies are indiscriminately amplified across all samples. To address this issue, we propose Dynamic Prototype Updating (DPU), a novel plug-and-play framework for multimodal OOD detection that accounts for intra-class variations. Our method dynamically updates class center representations for each class by measuring the variance of similar samples within each batch, enabling tailored adjustments. This approach allows us to intensify prediction discrepancies based on the updated class centers, thereby enhancing the model's robustness and generalization across different modalities. Extensive experiments on two tasks, five datasets, and nine base OOD algorithms demonstrate that DPU significantly improves OOD detection performances, setting a new state-of-the-art in multimodal OOD detection, including improvements up to 80% in Far-OOD detection.To improve accessibility and reproducibility, our code is released anonymously at https://anonymous.4open.science/r/CVPR-9177.
****
Towards Unbiased and Robust Spatio-Temporal Scene Graph Generation and Anticipation
Rohith Peddi, Saurabh Saurabh, Ayush Abhay Shrivastava, Parag Singla, Vibhav Gogate; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8648-8657
Spatio-Temporal Scene Graphs (STSGs) provide a concise and expressive representation of dynamic scenes by modeling objects and their evolving relationships over time. However, real-world visual relationships often exhibit a long-tailed distribution, causing existing methods for tasks like Video Scene Graph Generation (VidSGG) and Scene Graph Anticipation (SGA) to produce biased scene graphs. To this end, we propose ImparTail, a novel training framework that leverages loss masking and curriculum learning to mitigate bias in the generation and anticipation of spatio-temporal scene graphs. Unlike prior methods that add extra architectural components to learn unbiased estimators, we propose an impartial training objective that reduces the dominance of head classes during learning and focuses on underrepresented tail relationships. Our curriculum-driven mask generation strategy further empowers the model to adaptively adjust its bias mitigation strategy over time, enabling more balanced and robust estimations. To thoroughly assess performance under various distribution shifts, we also introduce two new tasks--Robust Spatio-Temporal Scene Graph Generation and Robust Scene Graph Anticipation--offering a challenging benchmark for evaluating the resilience of STSG models. Extensive experiments on the Action Genome dataset demonstrate the superior unbiased performance and robustness of our method compared to existing baselines.
****
Spatial Transport Optimization by Repositioning Attention Map for Training-Free Text-to-Image Synthesis
Woojung Han, Yeonkyung Lee, Chanyoung Kim, Kwanghyun Park, Seong Jae Hwang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18401-18410
Diffusion-based text-to-image (T2I) models have recently excelled in high-qualit

y image generation, particularly in a training-free manner, enabling cost-effect ive adaptability and generalization across diverse tasks. However, while the exi sting methods have been continuously focusing on several challenges such as "mis sing objects" and "mismatched attributes," another critical issue of "mislocated objects" remains where generated spatial positions fail to align with text prom pts. Surprisingly, ensuring such seemingly basic functionality remains challengi ng in popular T2I models due to the inherent difficulty of imposing explicit spa tial guidance via text forms. To address this, we propose STORM (Spatial Transpo rt Optimization by Repositioning Attention Map), a novel training-free approach for spatially coherent T2I synthesis. STORM employs Spatial Transport Optimizati on (STO), rooted in optimal transport theory, to dynamically adjust object atten tion maps for precise spatial adherence, supported by a Spatial Transport (ST) C ost function that enhances spatial understanding. Our analysis shows that integr ating spatial awareness is most effective in the early denoising stages, while l ater phases refine details. Extensive experiments demonstrate that STORM surpass es existing methods, effectively mitigating mislocated objects while improving m issing and mismatched attributes, setting a new benchmark for spatial alignment in T2I synthesis.

****

From Alexnet to Transformers: Measuring the Non-linearity of Deep Neural Network s with Affine Optimal Transport

Quentin Bouniot, Ievgen Redko, Anton Mallasto, Charlotte Laclau, Oliver Struckme ier, Karol Arndt, Markus Heinonen, Ville Kyrki, Samuel Kaski; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25250-2526 0

In the last decade, we have witnessed the introduction of several novel deep neu ral network (DNN) architectures exhibiting ever-increasing performance across di verse tasks. Explaining the upward trend of their performance, however, remains difficult as different DNN architectures of comparable depth and width -- common factors associated with their expressive power -- may exhibit a drastically dif ferent performance even when trained on the same dataset. In this paper, we intr oduce the concept of the non-linearity signature of DNN, the first theoretically sound solution for approximately measuring the non-linearity of deep neural net works. Built upon a score derived from closed-form optimal transport mappings, t his signature provides a better understanding of the inner workings of a wide ra nge of DNN architectures and learning paradigms, with a particular emphasis on t he computer vision task. We provide extensive experimental results that highligh t the practical usefulness of the proposed non-linearity signature and its poten tial for long-reaching implications.

****

Prompt2Perturb (P2P): Text-Guided Diffusion-Based Adversarial Attack on Breast U ltrasound Images

Yasamin Medghalchi, Moein Heidari, Clayton Allard, Leonid Sigal, Ilker Hacihalil oglu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVP R), 2025, pp. 28564-28574

Deep neural networks (DNNs) offer significant promise for improving breast canc er diagnosis in medical imaging. However, these models are highly susceptible to adversarial attacks--small, imperceptible changes that can mislead classifiers- -raising critical concerns about their reliability and security. Traditional att ack methods typically either require substantial extra data for malicious model pre-training, or involve a fixed norm perturbation budget, which does not align with human perception of these alterations. In medical imaging, however, this is often unfeasible due to the limited availability of datasets. Building on recen t advancements in learnable prompts, we propose Prompt2Perturb (P2P), a novel la nguage-guided semantic attack method capable of generating meaningful perturbati ons driven by text instructions. During the prompt learning phase, our approach leverages learnable prompts within the text encoder to create subtle, yet impact ful, perturbations that remain imperceptible while guiding the model towards tar geted outcomes.In contrast to current prompt learning-based approaches, our P2P stands out by directly updating text embeddings, avoiding the need for retrainin

g diffusion models or using large pre-trained models which is typically infeasib
le in medical domain. Further, we leverage the finding that optimizing only the
early reverse diffusion steps boosts efficiency while ensuring that the generate
d adversarial examples incorporate subtle low-frequency noise, thus preserving u
ltrasound image quality without introducing noticeable artifacts. We show that o
ur method outperforms state-of-the-art attack techniques across three breast ult
rasound datasets. Moreover, the generated images are both more natural in appear
ance and more effective compared to existing adversarial attacks.
****

Boosting Domain Incremental Learning: Selecting the Optimal Parameters is All Yo
u Need
Qiang Wang, Xiang Song, Yuhang He, Jizhou Han, Chenhao Ding, Xinyuan Gao, Yihong
 Gong; Proceedings of the Computer Vision and Pattern Recognition Conference (CV
PR), 2025, pp. 4839-4849
Deep neural networks (DNNs) often underperform in real-world, dynamic settings w
here data distributions change over time. Domain Incremental Learning (DIL) offe
rs a solution by enabling continuous model adaptation, with Parameter-Isolation
DIL (PIDIL) emerging as a promising paradigm to reduce knowledge conflicts. Howe
ver, existing PIDIL methods struggle with parameter selection accuracy, especial
ly as the number of domains and corresponding classes grows. To address this, we
 propose SOYO, a lightweight framework that improves domain selection in PIDIL.
SOYO introduces a Gaussian Mixture Compressor (GMC) and Domain Feature Resampler
 (DFR) to store and balance prior domain data efficiently, while a Multi-level D
omain Feature Fusion Network (MDFN) enhances domain feature extraction. Our fram
ework supports multiple Parameter-Efficient Fine-Tuning (PEFT) methods and is va
lidated across tasks such as image classification, object detection, and speech
enhancement. Experimental results on six benchmarks demonstrate SOYO's consisten
t superiority over existing baselines, showcasing its robustness and adaptabilit
y in complex, evolving environments.
****

COAP: Memory-Efficient Training with Correlation-Aware Gradient Projection
Jinqi Xiao, Shen Sang, Tiancheng Zhi, Jing Liu, Qing Yan, Linjie Luo, Bo Yuan; P
roceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 202
5, pp. 30116-30126
Training large-scale neural networks in vision, and multimodal domains demands s
ubstantial memory resources, primarily due to the storage of optimizer states. W
hile LoRA, a popular parameter-efficient method, reduces memory usage, it often
suffers from suboptimal performance due to the constraints of low-rank updates.
Low-rank gradient projection methods (e.g., GaLore, Flora) reduce optimizer memo
ry by projecting gradients and moment estimates into low-rank spaces via singula
r value decomposition or random projection. However, they fail to account for in
ter-projection correlation, causing performance degradation, and their projectio
n strategies often incur high computational costs. In this paper, we present COA
P (Correlation-Aware Gradient Projection), a memory-efficient method that minimi
zes computational overhead while maintaining training performance. Evaluated acr
oss various vision, language, and multimodal tasks, COAP outperforms existing me
thods in both training speed and model performance. For LLaMA-1B, it reduces opt
imizer memory by 61% with only 2% additional time cost, achieving the same PPL a
s AdamW. With 8-bit quantization, COAP cuts optimizer memory by 81% and achieves
 4x speedup over GaLore for LLaVA-v1.5-7B fine-tuning, while delivering higher a
ccuracy.
****

Perceptual Inductive Bias Is What You Need Before Contrastive Learning
Junru Zhao, Tianqin Li, Dunhan Jiang, Shenghao Wu, Alan Ramirez, Tai Sing Lee; P
roceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 202
5, pp. 9621-9630
David Marr's seminal theory of human perception stipulates that visual processin
g is a multi-stage process, prioritizing the derivation of boundary and surface
properties before forming semantic object representations. In contrast, contrast
ive representation learning frameworks typically bypass this explicit multi-stag

e approach, defining their objective as the direct learning of a semantic repres entation space for objects. While effective in general contexts, this approach s acrifices the inductive biases of vision, leading to slower convergence speed an d learning shortcut resulting in texture bias. In this work, we demonstrate that leveraging Marr's multi-stage theory--by first constructing boundary and surfac e-level representations using perceptual constructs from early visual processing stages and subsequently training for object semantics--leads to 2x faster conve rgence on ResNet18, improved final representations on semantic segmentation, dep th estimation, and object recognition, and enhanced robustness and out-of-distri bution capability. Together, we propose a pretraining stage before the general c ontrastive representation pretraining to further enhance the final representatio n quality and reduce the overall convergence time via inductive bias from human vision systems.
****

FaceBench: A Multi-View Multi-Level Facial Attribute VQA Dataset for Benchmarkin g Face Perception MLLMs

Xiaoqin Wang, Xusen Ma, Xianxu Hou, Meidan Ding, Yudong Li, Junliang Chen, Wenti ng Chen, Xiaoyang Peng, Linlin Shen; Proceedings of the Computer Vision and Patt ern Recognition Conference (CVPR), 2025, pp. 9154-9164

Multimodal large language models (MLLMs) have demonstrated remarkable capabiliti es in various tasks. However, effectively evaluating these MLLMs on face percept ion remains largely unexplored. To address this gap, we introduce FaceBench, a d ataset featuring hierarchical multi-view and multi-level attributes specifically designed to assess the comprehensive face perception abilities of MLLMs. Initia lly, we construct a hierarchical facial attribute structure, which encompasses f ive views with up to three levels of attributes, totaling over 210 attributes an d 700 attribute values. Based on the structure, the proposed FaceBench consists of 49,919 visual question-answering (VQA) pairs for evaluation and 23,841 pairs for fine-tuning. Moreover, we further develop a robust face perception MLLM base line, Face-LLaVA, by training with our proposed face VQA data. Extensive experim ents on various mainstream MLLMs and Face-LLaVA are conducted to test their face perception ability, with results also compared against human performance. The r esults reveal that, the existing MLLMs are far from satisfactory in understandin g the fine-grained facial attributes, while our Face-LLaVA significantly outperf orms existing open-source models with a small amount of training data and is com parable to commercial ones like GPT-4o and Gemini. The dataset will be released at https://github.com/CVI-SZU/FaceBench.
****

EffiDec3D: An Optimized Decoder for High-Performance and Efficient 3D Medical Im age Segmentation

Md Mostafijur Rahman, Radu Marculescu; Proceedings of the Computer Vision and Pa ttern Recognition Conference (CVPR), 2025, pp. 10435-10444

Recent 3D deep networks such as SwinUNETR, SwinUNETRv2, and 3D UX-Net have shown promising performance by leveraging self-attention and large-kernel convolution s to capture the volumetric context. However, their substantial computational re quirements limit their use in real-time and resource-constrained environments. T he high #FLOPs and #Params in these networks stem largely from complex decoder d esigns with high-resolution layers and excessive channel counts. In this paper, we propose EffiDec3D, an optimized 3D decoder that employs a channel reduction s trategy across all decoder stages, which sets the number of channels to the mini mum needed for accurate feature representation. Additionally, EffiDec3D removes the high-resolution layers when their contribution to segmentation quality is mi nimal. Our optimized EffiDec3D decoder achieves a 96.4% reduction in #Params and a 93.0% reduction in #FLOPs compared to the decoder of original 3D UX-Net. Simi larly, for SwinUNETR and SwinUNETRv2 (which share an identical decoder), we obse rve reductions of 94.9% in #Params and 86.2% in #FLOPs. Our extensive experiment s on 12 different medical imaging tasks confirm that EffiDec3D not only signific antly reduces the computational demands, but also maintains a performance level comparable to original models, thus establishing a new standard for efficient 3D medical image segmentation. Our implementation is available at https://github.c

om/SLDGroup/EffiDec3D.
****
Exploring Historical Information for RGBE Visual Tracking with Mamba
Chuanyu Sun, Jiqing Zhang, Yang Wang, Huilin Ge, Qianchen Xia, Baocai Yin, Xin Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6500-6509
Combining the advantages of conventional and event cameras for robust visual tracing has drawn extensive interest. However, existing tracking approaches heavily engage in complex cross-modal fusion modules, leading to higher computational complexity and training challenges. Besides, these methods generally ignore the effective integration of historical information, which is crucial to grasping the change in the target's appearance and motion trends. Given the recent advancements in Mamba's long-range modeling and linear complexity, we explore its potential in addressing the above issues in RGBE tracking tasks. Specifically, we first propose an efficient fusion module based on Mamba, which utilizes a simple gate-based interaction scheme to achieve effective modality-selective fusion. This module can be seamlessly integrated into the encoding layer of prevalent Transformer-based backbones. Moreover, we further present a novel historical decoder that leverages Mamba's advanced long sequence modeling to effectively capture the target appearance changes with autoregressive queries. Extensive experiments show that our proposed approach achieves state-of-the-art performance on multiple challenging short-term and long-term RGBE benchmarks. Besides, the effectiveness of each key Mamba-based component of our approach is evidenced by our thorough ablation study.
****
Gyro-based Neural Single Image Deblurring
Heemin Yang, Jaesung Rim, Seungyong Lee, Seung-Hwan Baek, Sunghyun Cho; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23111-23120
In this paper, we present GyroDeblurNet, a novel single-image deblurring method that utilizes a gyro sensor to resolve the ill-posedness of image deblurring. The gyro sensor provides valuable information about camera motion that can improve deblurring quality. However, exploiting real-world gyro data is challenging due to errors from various sources. To handle these errors, GyroDeblurNet is equipped with two novel neural network blocks: a gyro refinement block and a gyro deblurring block. The gyro refinement block refines the erroneous gyro data using the blur information from the input image. The gyro deblurring block removes blur from the input image using the refined gyro data and further compensates for gyro error by leveraging the blur information from the input image. For training a neural network with erroneous gyro data, we propose a training strategy based on the curriculum learning. We also introduce a novel gyro data embedding scheme to represent real-world intricate camera shakes. Finally, we present both synthetic and real-world datasets for training and evaluating gyro-based single image deblurring. Our experiments demonstrate that our approach achieves state-of-the-art deblurring quality by effectively utilizing erroneous gyro data.
****
ArtiScene: Language-Driven Artistic 3D Scene Generation Through Image Intermediary
Zeqi Gu, Yin Cui, Zhaoshuo Li, Fangyin Wei, Yunhao Ge, Jinwei Gu, Ming-Yu Liu, Abe Davis, Yifan Ding; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2891-2901
Designing 3D scenes is traditionally a challenging and laborious task that demands both artistic expertise and proficiency with complex software. Recent advances in text-to-3D generation have greatly simplified this process by letting users create scenes based on simple text descriptions. However, as these methods generally require extra training or in-context learning, their performance is often hindered by the limited availability of high-quality 3D data. In contrast, modern text-to-image models learned from web-scale images can generate scenes with diverse, reliable spatial layouts and consistent, visually appealing styles. Our key insight is that instead of learning directly from 3D scenes, we can leverage

generated 2D images as an intermediary to guide 3D synthesis. In light of this, we introduce ArtiScene, a training-free automated pipeline for scene design that integrates the flexibility of free-form text-to-image generation with the diversity and reliability of 2D intermediary layouts. We generate the 2D intermediary image from a scene description, extract object shapes and appearances, create 3D models, and assemble them into the final scene with geometry, position, and pose extracted from the same image. Being generalizable to a wide range of scenes and styles, ArtiScene is shown to outperform state-of-the-art benchmarks by a large margin in layout and aesthetic quality by quantitative metrics. It also averages a 74.89% winning rate in extensive user studies and 95.07% in GPT evaluation.

****

MobileH2R: Learning Generalizable Human to Mobile Robot Handover Exclusively from Scalable and Diverse Synthetic Data

Zifan Wang, Ziqing Chen, Junyu Chen, Jilong Wang, Yuxin Yang, Yunze Liu, Xueyi Liu, He Wang, Li Yi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17315-17325

This paper introduces MobileH2R, a framework for learning generalizable vision-based human-to-mobile-robot (H2MR) handover skills. Unlike traditional fixed-base handovers, this task requires a mobile robot to reliably receive objects in a large workspace enabled by its mobility. Our key insight is that generalizable handover skills can be developed in simulators using high-quality synthetic data, without the need for real-world demonstrations. To achieve this, we propose a scalable pipeline for generating diverse synthetic full-body human motion data, an automated method for creating safe and imitation-friendly demonstrations, and an efficient 4D imitation learning method for distilling large-scale demonstrations into closed-loop policies with base-arm coordination. Experimental evaluations in both simulators and the real world show significant improvements (at least +15% success rate) over baseline methods in all cases. Experiments also validate that large-scale and diverse synthetic data greatly enhances robot learning, highlighting our scalable framework.

****

Improving Sound Source Localization with Joint Slot Attention on Image and Audio

Inho Kim, Youngkil Song, Jicheol Park, Won Hwa Kim, Suha Kwak; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3121-3130

Sound source localization (SSL) is the task of locating the source of sound within an image. Due to the lack of localization labels, the de facto standard in SSL has been to represent an image and audio as a single embedding vector each, and use them to learn SSL via contrastive learning. To this end, previous work samples one of local image features as the image embedding and aggregates all local audio features to obtain the audio embedding, which is far from optimal due to the presence of noise and background irrelevant to the actual target in the input.We present a novel SSL method that addresses this chronic issue by joint slot attention on image and audio. To be specific, two slots competitively attend image and audio features to decompose them into target and off-target representations, and only target representations of image and audio are used for contrastive learning. Also, we introduce cross-modal attention matching to further align local features of image and audio. Our method achieved the best in almost all settings on three public benchmarks for SSL, and substantially outperformed all the prior work in cross-modal retrieval.

****

Improved Monocular Depth Prediction Using Distance Transform Over Pre-semantic Contours with Self-supervised Neural Networks

Marwane Hariat, Antoine Manzanera, David Filliat; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21868-21879

Monocular depth estimation (MDE) with self-supervised training approaches struggles in low-texture areas, where photometric losses may lead to ambiguous depth predictions. To address this, we propose a novel technique that enhances spatial information by applying a distance transform over pre-semantic contours, augmenting discriminative power in low texture regions. Our approach jointly estimates

pre-semantic contours, depth and ego-motion. The pre-semantic contours are leveraged to produce new input images, with variance augmented by the distance transform in uniform areas. This approach results in more effective loss functions, enhancing the training process for depth and ego-motion. We demonstrate theoretically that the distance transform is the optimal variance-augmenting technique in this context. Through extensive experiments on KITTI and Cityscapes, our model demonstrates robust performance, surpassing conventional self-supervised methods in MDE.

****

Feature-Preserving Mesh Decimation for Normal Integration

Moritz Heep, Sven Behnke, Eduard Zell; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5783-5792

Normal integration reconstructs 3D surfaces from normal maps obtained, e.g. by photometric stereo. These normal maps capture surface details down to the pixel level but require large computational resources for integration at high resolutions. In this work, we replace the dense pixel grid with a sparse anisotropic triangle mesh prior to normal integration. We adapt the triangle mesh to the local geometry in the case of complex surface structures and remove oversampling from flat featureless regions. For high-resolution images, the resulting compression reduces normal integration runtimes from hours to minutes while maintaining high surface accuracy. Our main contribution is the derivation of the well-known quadric error measure from mesh decimation for screen space applications and its combination with optimal Delaunay triangulation.

****

Is this Generated Person Existed in Real-world? Fine-grained Detecting and Calibrating Abnormal Human-body

Zeqing Wang, Qingyang Ma, Wentao Wan, Haojie Li, Keze Wang, Yonghong Tian; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21226-21237

Recent improvements in visual synthesis have significantly enhanced the depiction of generated human photos, which are pivotal due to their wide applicability and demand. Nonetheless, the existing text-to-image or text-to-video models often generate low-quality human photos that might differ considerably from real-world body structures, referred to as "abnormal human bodies". Such abnormalities, typically deemed unacceptable, pose considerable challenges in the detection and repair of them within human photos. These challenges require precise abnormality recognition capabilities, which entail pinpointing both the location and the abnormality type. Intuitively, Visual Language Models (VLMs) that have obtained remarkable performance on various visual tasks are quite suitable for this task. However, their performance on abnormality detection in human photos is quite poor.Hence, it is quite important to highlight this task for the research community. In this paper, we first introduce a simple yet challenging task, i.e., Fine-grained Human-body Abnormality Detection (FHAD), and construct two high-quality datasets for evaluation. Then, we propose a meticulous framework, named HumanCalibrator, which identifies and repairs abnormalities in human body structures while preserving the other content. Experiments indicate that our HumanCalibrator achieves high accuracy in abnormality detection and accomplishes an increase in visual comparisons while preserving the other visual content.

****

PERSE: Personalized 3D Generative Avatars from A Single Portrait

Hyunsoo Cha, Inhee Lee, Hanbyul Joo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15953-15962

We present PERSE, a method for building a personalized 3D generative avatar from a reference portrait. Our avatar enables facial attribute editing in a continuous and disentangled latent space to control each facial attribute, while preserving the individual's identity. To achieve this, our method begins by synthesizing large-scale synthetic 2D video datasets, where each video contains consistent changes in facial expression and viewpoint, along with variations in a specific facial attribute from the original input. We propose a novel pipeline to produce high-quality, photorealistic 2D videos with facial attribute editing. Leveragin

g this synthetic attribute dataset, we present a personalized avatar creation method based on 3D Gaussian Splatting, learning a continuous and disentangled latent space for intuitive facial attribute manipulation. To enforce smooth transitions in this latent space, we introduce a latent space regularization technique by using interpolated 2D faces as supervision. Compared to previous approaches, we demonstrate that PERSE generates high-quality avatars with interpolated attributes while preserving the identity of the reference individual.
****

Automated Generation of Challenging Multiple-Choice Questions for Vision Language Model Evaluation
Yuhui Zhang, Yuchang Su, Yiming Liu, Xiaohan Wang, James Burgess, Elaine Sui, Chenyu Wang, Josiah Aklilu, Alejandro Lozano, Anjiang Wei, Ludwig Schmidt, Serena Yeung-Levy; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29580-29590
The rapid development of vision language models (VLMs) demands rigorous and reliable evaluation. However, current visual question answering (VQA) benchmarks often depend on open-ended questions, making accurate evaluation difficult due to the variability in natural language responses. To address this, we introduce AutoConverter, an agentic framework that automatically converts these open-ended questions into multiple-choice format, enabling objective evaluation while reducing the costly multiple-choice question creation process. Our experiments demonstrate that AutoConverter can generate correct and challenging multiple-choice questions, with VLMs demonstrating consistently similar or lower accuracy on these questions compared to human-created ones. Using AutoConverter, we construct VMCBench, a benchmark created by transforming 20 existing VQA datasets into a unified multiple-choice format, totaling 9,018 questions. We comprehensively evaluate 33 state-of-the-art VLMs on VMCBench, setting a new standard for scalable, consistent, and reproducible VLM evaluation.
****

DexHandDiff: Interaction-aware Diffusion Planning for Adaptive Dexterous Manipulation
Zhixuan Liang, Yao Mu, Yixiao Wang, Tianxing Chen, Wenqi Shao, Wei Zhan, Masayoshi Tomizuka, Ping Luo, Mingyu Ding; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1745-1755
Dexterous manipulation with contact-rich interactions is crucial for advanced robotics. While recent diffusion-based planning approaches show promise for simple manipulation tasks, they often produce unrealistic ghost states (e.g., the object automatically moves without hand contact) or lack adaptability when handling complex sequential interactions. In this work, we introduce DexHandDiff, an interaction-aware diffusion planning framework for adaptive dexterous manipulation. DexHandDiff models joint state-action dynamics through a dual-phase diffusion process which consists of pre-interaction contact alignment and post-contact goal-directed control, enabling goal-adaptive generalizable dexterous manipulation. Additionally, we incorporate dynamics model-based dual guidance and leverage large language models for automated guidance function generation, enhancing generalizability for physical interactions and facilitating diverse goal adaptation through language cues. Experiments on physical interaction tasks such as door opening, pen and block re-orientation, object relocation, and hammer striking demonstrate DexHandDiff's effectiveness on goals outside training distributions, achieving over twice the average success rate (59.2% vs. 29.5%) compared to existing methods. Our framework achieves an average of 70.7% success rate on goal adaptive dexterous tasks, highlighting its robustness and flexibility in contact-rich manipulation.
****

VerbDiff: Text-Only Diffusion Models with Enhanced Interaction Awareness
SeungJu Cha, Kwanyoung Lee, Ye-Chan Kim, Hyunwoo Oh, Dong-Jin Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8041-8050
Recent large-scale text-to-image diffusion models generate photorealistic images but often struggle to accurately depict interactions between humans and objects

due to their limited ability to differentiate various interaction words.In this work, we propose VerbDiff to address the challenge of capturing nuanced interactions within text-to-image diffusion models. VerbDiff is a novel text-to-image generation model that weakens the bias between interaction words and objects, enhancing the understanding of interactions. Specifically, we disentangle various interaction words from frequency-based anchor words and leverage localized interaction regions from generated images to help the model better capture semantics in distinctive words without extra conditions. Our approach enables the model to accurately understand the intended interaction between humans and objects, producing high-quality images with accurate interactions aligned with specified verbs. Extensive experiments on the HICO-DET dataset demonstrate the effectiveness of our method compared to previous approaches.

****

ROLL: Robust Noisy Pseudo-label Learning for Multi-View Clustering with Noisy Correspondence

Yuan Sun, Yongxiang Li, Zhenwen Ren, Guiduo Duan, Dezhong Peng, Peng Hu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30732-30741

Multi-view clustering (MVC) aims to exploit complementary information from diverse views to enhance clustering performance. Since pseudo-labels can provide additional semantic information, many MVC methods have been proposed to guide unsupervised multi-view learning through pseudo-labels. These methods implicitly assume that the predicted pseudo-labels are predicted correctly. However, due to the challenges in training a flawless unsupervised model, this assumption can be easily violated, thereby leading to the Noisy Pseudo-label Problem (NPP). Moreover, these existing approaches typically rely on the assumption of perfect cross-view alignment. In practice, it is frequently compromised due to noise or sensor differences, thereby resulting in the Noisy Correspondence Problem (NCP). Based on the above observations, we reveal and study unsupervised multi-view learning under NPP and NCP. To this end, we propose Robust Noisy Pseudo-label Learning (ROLL) to prevent the overfitting problem caused by both NPP and NCP. Specifically, we first adopt traditional contrastive learning to warm up the model, thereby generating the pseudo-labels in a self-supervised manner. Afterward, we propose noise-tolerance pseudo-label learning to deal with the noise in the predicted pseudo-labels, thereby embracing the robustness against NPP. To further mitigate the overfitting problem, we present robust multi-view contrastive learning to mitigate the negative impact of NCP. Extensive experiments on five multi-view datasets demonstrate the superior clustering performance of our ROLL compared to 11 state-of-the-art methods.

****

Towards In-the-wild 3D Plane Reconstruction from a Single Image

Jiachen Liu, Rui Yu, Sili Chen, Sharon X. Huang, Hengkai Guo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27027-27037

3D plane reconstruction from a single image is a crucial yet challenging topic in 3D computer vision. Previous state-of-the-art (SOTA) methods have focused on training their system on a single dataset from either indoor or outdoor domain, limiting their generalizability across diverse testing data. In this work, we introduce a novel framework dubbed ZeroPlane, a Transformer-based model targeting zero-shot 3D plane detection and reconstruction from a single image, over diverse domains and environments. To enable data-driving models on multiple domains, we have curated a large-scale (over 14 datasets and 560,000 images), high-resolution, densely-annotated planar benchmark from various indoor and outdoor scenes. To address the challenge of achieving desirable planar geometry on multi-dataset training, we propose to disentangle the representation of plane normal and offset, and employ an exemplar-guided, classification-then-regression paradigm to learn plane and offset respectively. Additionally, we employ advanced backbones as image encoder, and present an effective pixel-geometry-enhanced plane embedding module to further facilitate planar reconstruction. Extensive experiments across multiple zero-shot evaluation datasets have demonstrated that our approach sign

ificantly outperforms previous methods on both reconstruction accuracy and generalizability, especially over in-the-wild data. We will release all of the labeled data, code and models upon the acceptance of this paper. Our code and data are available at: https://github.com/jcliu0428/ZeroPlane.
****
Memories of Forgotten Concepts
Matan Rusanovsky, Shimon Malnick, Amir Jevnisek, Ohad Fried, Shai Avidan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2966-2975
Diffusion models dominate the space of text-to-image generation, yet they may produce undesirable outputs, including explicit content or private data. To mitigate this, concept ablation techniques have been explored to limit the generation of certain concepts.In this paper, we reveal that the erased concept information persists in the model and that erased concept images can be generated using the right latent. Utilizing inversion methods, we show that there exist latent seeds capable of generating high quality images of erased concepts.Moreover, we show that these latents have likelihoods that overlap with those of images outside the erased concept.We extend this to demonstrate that for every image from the erased concept set, we can generate many seeds that generate the erased concept.Given the vast space of latents capable of generating ablated concept images, our results suggest that fully erasing concept information may be intractable, highlighting possible vulnerabilities in current concept ablation techniques.
****
Dynamic Stereotype Theory Induced Micro-expression Recognition with Oriented Deformation
Bohao Zhang, Xuejiao Wang, Changbo Wang, Gaoqi He; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10701-10711
Micro-expression recognition (MER) aims to uncover genuine emotions and underlying psychological states. However, existing MER methods struggle with three main challenges. 1) Scarcity of micro-expression samples. 2) Difficulty in modeling nearly imperceptible facial movements. 3) Reliance on apex frame annotations. To address these issues, we propose a Self-supervised Oriented Deformation model for Apex-free Micro-expression Recognition (SODA4MER). Our approach enhances local deformation perception using muscle-group priors and amplifies subtle features through Dynamic Stereotype Theory (DST) based enhancement, while contrastive learning eliminates the need for manual apex annotations. Specifically, the Oriented deformation estimator of SODA4MER is first pre-trained in a self-supervised manner. Secondly, a Gated Temporal Variance Gaussian model (GTVG) is introduced to adaptively integrate facial muscle-group priors, enhancing local deformation perception and mitigating noise from head movements. Then, contrastive learning is employed to achieve apex detection by identifying the frame with the most significant local deformation. Finally, guided by DST, we introduced a feature enhancement strategy that models the temporal dynamics of local deformation in the activation and decay phases, leading to richer deformation features. Our rigorous experiments confirm the competitive performance and practical applicability of SODA4MER.
****
PQPP: A Joint Benchmark for Text-to-Image Prompt and Query Performance Prediction
Eduard Poesina, Adriana Valentina Costache, Adrian-Gabriel Chifu, Josiane Mothe, Radu Tudor Ionescu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28651-28661
Text-to-image generation has recently emerged as a viable alternative to text-to-image retrieval, driven by the visually impressive results of generative diffusion models. Although query performance prediction is an active research topic in information retrieval, to the best of our knowledge, there is no prior study that analyzes the difficulty of queries (referred to as prompts) in text-to-image generation, based on human judgments. To this end, we introduce the first dataset of prompts which are manually annotated in terms of image generation performance. Additionally, we extend these evaluations to text-to-image retrieval by coll

ecting manual annotations that represent retrieval performance. We thus establis h the first joint benchmark for prompt and query performance prediction (PQPP) a cross both tasks, comprising over 10K queries. Our benchmark enables (i) the com parative assessment of prompt/query difficulty in both image generation and imag e retrieval, and (ii) the evaluation of prompt/query performance predictors addr essing both generation and retrieval. We evaluate several pre- and post-generati on/retrieval performance predictors, thus providing competitive baselines for fu ture research. Our benchmark and code are publicly available at https://github.c om/Eduard6421/PQPP.
****

CheXwhatsApp: A Dataset for Exploring Challenges in the Diagnosis of Chest X-ray s through Mobile Devices
Mariamma Antony, Rajiv Porana, Sahil M Lathiya, Siva Teja Kakileti, Chiranjib Bh attacharyya; Proceedings of the Computer Vision and Pattern Recognition Conferen ce (CVPR), 2025, pp. 25887-25896
Mobile health (mHealth) has emerged as a transformative solution to enhance heal thcare accessibility and affordability, particularly in resource-constrained reg ions and low-to-middle-income countries.mHealth leverages mobile platforms to im prove healthcare accessibility, addressing radiologist shortages in low-resource settings by enabling remote diagnosis and consultation through mobile devices. Mobile phones allow healthcare workers to transmit radiographic images, such as chest X-rays (CXR), to specialists or AI-driven models for interpretation. Howev er, AI-based diagnosis using CXR images shared via apps like WhatsApp suffers fr om reduced predictability and explainability due to compression artifacts, and t here is a lack of datasets to systematically study these challenges. To address this, we introduce CheXwhatsApp, a dataset of 141,804 paired original and WhatsA pp-compressed CXR images. We present a benchmarking study which shows the datase t improves prediction stability and explainability of state-of-the-art models by up to 80%, while also enhancing localization performance. CheXwhatsApp is open-sourced to support advancements in mHealth applications for CXR analysis.
****

PSBD: Prediction Shift Uncertainty Unlocks Backdoor Detection
Wei Li, Pin-Yu Chen, Sijia Liu, Ren Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10255-10264
Deep neural networks are susceptible to backdoor attacks, where adversaries mani pulate model predictions by inserting malicious samples into the training data. Currently, there is still a significant challenge in identifying suspicious trai ning data to unveil potential backdoor samples. In this paper, we propose a nove l method, Prediction Shift Backdoor Detection (PSBD), leveraging an uncertainty-based approach requiring minimal unlabeled clean validation data. PSBD is motiva ted by an intriguing Prediction Shift (PS) phenomenon, where poisoned models' pr edictions on clean data often shift away from true labels towards certain other labels with dropout applied during inference, while backdoor samples exhibit les s PS. We hypothesize PS results from the neuron bias effect, making neurons favo r features of certain classes. PSBD identifies backdoor training samples by comp uting the Prediction Shift Uncertainty (PSU), the variance in probability values when dropout layers are toggled on and off during model inference. Extensive ex periments have been conducted to verify the effectiveness and efficiency of PSBD , which achieves state-of-the-art results among mainstream detection methods. Th e code is available at https://github.com/WL-619/PSBD.
****

Degradation-Aware Feature Perturbation for All-in-One Image Restoration
Xiangpeng Tian, Xiangyu Liao, Xiao Liu, Meng Li, Chao Ren; Proceedings of the Co mputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28165-28175
All-in-one image restoration aims to recover clear images from various degradati on types and levels with a unified model. Nonetheless, the significant variation s among degradation types present challenges for training a universal model, oft en resulting in task interference, where the gradient update directions of diffe rent tasks may diverge due to shared parameters. To address this issue, motivate d by the routing strategy, we propose DFPIR, a novel all-in-one image restorer t

hat introduces Degradation-aware Feature Perturbations(DFP) to adjust the featur e space to align with the unified parameter space. In this paper, the feature pe rturbations primarily include channel-wise perturbations and attention-wise pert urbations. Specifically, channel-wise perturbations are implemented by shuffling the channels in high-dimensional space guided by degradation types, while atten tion-wise perturbations are achieved through selective masking in the attention space. To achieve these goals, we propose a Degradation-Guided Perturbation Bloc k (DGPB) to implement these two functions, positioned between the encoding and d ecoding stages of the encoder-decoder architecture. Extensive experimental resul ts demonstrate that DFPIR achieves state-of-the-art performance on several all-i n-one image restoration tasks including image denoising, image dehazing, image d eraining, motion deblurring, and low-light image enhancement. Our codes are avai lable at https://github.com/TxpHome/DFPIR.

****

ACL: Activating Capability of Linear Attention for Image Restoration

Yubin Gu, Yuan Meng, Jiayi Ji, Xiaoshuai Sun; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17913-17923

Image restoration (IR), a key area in computer vision, has entered a new era wit h deep learning. Recent research has shifted toward Selective State Space Models (Mamba) to overcome CNNs' limited receptive fields and Transformers' computatio nal inefficiency. However, due to Mamba's inherent one-dimensional scanning limi tations, recent approaches have introduced multi-directional scanning to bolster inter-sequence correlations. Despite these enhancements, these methods still st ruggle with managing local pixel correlations across various directions. Moreove r, the recursive computation in Mamba's SSM leads to reduced efficiency. To reso lve these issues, we exploit the mathematical congruences between linear attenti on and SSM within Mamba to propose a novel model, ACL, which leverages news desi gns to Activate the Capability of Linear attention for IR. ACL integrates linear attention blocks instead of SSM within Mamba, serving as the core component of encoders/decoders, and aims to preserve a global perspective while boosting comp utational efficiency. Furthermore, we have designed a simple yet robust local en hancement module with multi-scale dilated convolutions to extract both coarse an d fine features to improve local detail recovery. Experimental results confirm t hat our ACL model excels in classical IR tasks such as de-raining and de-blurrin g, while maintaining relatively low parameter counts and FLOPs.

****

GenDeg: Diffusion-based Degradation Synthesis for Generalizable All-In-One Image Restoration

Sudarshan Rajagopalan, Nithin Gopalakrishnan Nair, Jay N. Paranjape, Vishal M. P atel; Proceedings of the Computer Vision and Pattern Recognition Conference (CVP R), 2025, pp. 28144-28154

Deep learning-based models for All-In-One image Restoration (AIOR) have achieved significant advancements in recent years. However, their practical applicabilit y is limited by poor generalization to samples outside the training distribution . This limitation arises primarily from insufficient diversity in degradation va riations and scenes within existing datasets, resulting in inadequate representa tions of real-world scenarios. Additionally, capturing large-scale real-world pa ired data for degradations such as haze, low-light, and raindrops is often cumbe rsome and sometimes infeasible. In this paper, we leverage the generative capabi lities of latent diffusion models to synthesize high-quality degraded images fro m their clean counterparts. Specifically, we introduce GenDeg, a degradation and intensity-aware conditional diffusion model, capable of producing diverse degra dation patterns on clean images. Using GenDeg, we synthesize over 550k samples a cross six degradation types: haze, rain, snow, motion blur, low-light, and raind rops. These generated samples are integrated with existing datasets to form the GenDS dataset, comprising over 750k samples. Our experiments reveal that image r estoration models trained on GenDS dataset exhibit significant improvements in o ut-of-distribution performance as compared to when trained solely on existing da tasets. Furthermore, we provide comprehensive analyses on implications of diffus ion model-based synthetic degradations for AIOR.

****
Phoenix: A Motion-based Self-Reflection Framework for Fine-grained Robotic Action Correction
Wenke Xia, Ruoxuan Feng, Dong Wang, Di Hu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6981-6990
Building a generalizable self-correction system is crucial for robots to recover from failures. Despite advancements in Multimodal Large Language Models (MLLMs) that empower robots with semantic reflection ability for failure, translating semantic reflection into how to correct fine-grained robotic actions remains a significant challenge. To address this gap, we build the Phoenix framework, which leverages motion instruction as a bridge to connect high-level semantic reflection with low-level robotic action correction. In this motion-based self-reflection framework, we start with a dual-process motion adjustment mechanism with MLLMs to translate the semantic reflection into coarse-grained motion instruction adjustment. To leverage this motion instruction for guiding how to correct fine-grained robotic actions, a multi-task motion-conditioned diffusion policy is proposed to integrate visual observations for high-frequency robotic action correction. By combining these two models, we could shift the demand for generalization capability from the low-level manipulation policy to the MLLMs-driven motion adjustment model and facilitate precise, fine-grained robotic action correction. Utilizing this framework, we further develop a lifelong learning method to automatically improve the model's capability from interactions with dynamic environments. The experiments conducted in both the RoboMimic simulation and real-world scenarios prove the superior generalization and robustness of our framework across a variety of manipulation tasks.
****
The Power of Context: How Multimodality Improves Image Super-Resolution
Kangfu Mei, Hossein Talebi, Mojtaba Ardakani, Vishal M. Patel, Peyman Milanfar, Mauricio Delbracio; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23141-23152
Single-image super-resolution (SISR) remains challenging due to the inherent difficulty of recovering fine-grained details and preserving perceptual quality from low-resolution inputs. Existing methods often rely on limited image priors, leading to suboptimal results. We propose a novel approach that leverages the rich contextual information available in multiple modalities -- including depth, segmentation, edges, and text prompts -- to learn a powerful generative prior for SISR within a diffusion model framework. We introduce a flexible network architecture that effectively fuses multimodal information, accommodating an arbitrary number of input modalities without requiring significant modifications to the diffusion process. Crucially, we mitigate hallucinations, often introduced by text prompts, by using spatial information from other modalities to guide regional text-based conditioning. Each modality's guidance strength can also be controlled independently, allowing steering outputs toward different directions, such as increasing bokeh through depth or adjusting object prominence via segmentation. Extensive experiments demonstrate that our model surpasses state-of-the-art generative SISR methods, achieving superior visual quality and fidelity.
****
MARBLE: Material Recomposition and Blending in CLIP-Space
Ta Ying Cheng, Prafull Sharma, Mark Boss, Varun Jampani; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13061-13071
Editing materials of objects in images based on exemplar images is an active area of research in computer vision and graphics. We propose MARBLE, a method for performing material blending and recomposing fine-grained material properties by finding material embeddings in CLIP-space and using that to control pre-trained text-to-image models. We improve exemplar-based material editing by finding a block in the denoising UNet responsible for material attribution. Given two material exemplar-images, we find directions in the CLIP-space for blending the materials. Further, we can achieve parametric control over fine-grained material attributes such as roughness, metallic, transparency, and glow using a shallow network to predict the direction for the desired material attribute change. We perform

qualitative and quantitative analysis to demonstrate the efficacy of our proposed method. We also present the ability of our method to perform multiple edits in a single forward pass and applicability to painting.
****

Multirate Neural Image Compression with Adaptive Lattice Vector Quantization
Hao Xu, Xiaolin Wu, Xi Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7633-7642
Recent research has explored integrating lattice vector quantization (LVQ) into learned image compression models. Due to its more efficient Voronoi covering of vector space than scalar quantization (SQ), LVQ achieves better rate-distortion (R-D) performance than SQ, while still retaining the low complexity advantage of SQ. However, existing LVQ-based methods have two shortcomings: 1) lack of a multirate coding mode, hence incapable to operate at different rates; 2) the use of a fixed lattice basis, hence nonadaptive to changing source distributions. To overcome these shortcomings, we propose a novel adaptive LVQ method, which is the first among LVQ-based methods to achieve both rate and domain adaptations. By scaling the lattice basis vector, our method can adjust the density of lattice points to achieve various bit rate targets, achieving superior R-D performance to current SQ-based variable rate models. Additionally, by using a learned invertible linear transformation between two different input domains, we can reshape the predefined lattice cell to better represent the target domain, further improving the R-D performance. To our knowledge, this paper represents the first attempt to propose a unified solution for rate adaptation and domain adaptation through quantizer design.
****

EventFly: Event Camera Perception from Ground to the Sky
Lingdong Kong, Dongyue Lu, Xiang Xu, Lai Xing Ng, Wei Tsang Ooi, Benoit R. Cottereau; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1472-1484
Cross-platform adaptation in event-based dense perception is crucial for deploying event cameras across diverse settings, such as vehicles, drones, and quadrupeds, each with unique motion dynamics, viewpoints, and class distributions. In this work, we introduce EventFly, a framework for robust cross-platform adaptation in event camera perception. Our approach comprises three key components: i) Event Activation Prior (EAP), which identifies high-activation regions in the target domain to minimize prediction entropy, fostering confident, domain-adaptive predictions; ii) EventBlend, a data-mixing strategy that integrates source and target event voxel grids based on EAP-driven similarity and density maps, enhancing feature alignment; and iii) EventMatch, a dual-discriminator technique that aligns features from source, target, and blended domains for better domain-invariant learning. To holistically assess cross-platform adaptation abilities, we introduce EXPo, a large-scale benchmark with diverse samples across vehicle, drone, and quadruped platforms. Extensive experiments validate our effectiveness, demonstrating substantial gains over popular adaptation methods. We hope this work can pave the way for more adaptive, high-performing event perception across diverse and complex environments.
****

Detect Any Mirrors: Boosting Learning Reliability on Large-Scale Unlabeled Data with an Iterative Data Engine
Zhaohu Xing, Lihao Liu, Yijun Yang, Hongqiu Wang, Tian Ye, Sixiang Chen, Wenxue Li, Guang Liu, Lei Zhu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25476-25486
Mirror detection is a challenging task because a mirror's visual appearance varies depending on the reflected content. Due to limited annotated data, current methods failed to generalize well for detecting diverse mirror scenes. Semi-supervised learning with large-scale unlabeled data can improve generalization capabilities on mirror detection, but these methods often suffer from unreliable pseudo-labels due to distribution differences between labeled and unlabeled data, therefore affecting the learning process. To address this issue, we first collect a large-scale dataset of approximately 0.4 million mirror-related images from the

internet, significantly expanding the data scale for mirror detection. To effect ively exploit this unlabeled dataset, we propose the first semi-supervised frame work (namely an iterative data engine) consisting of four steps: (1) mirror dete ction model training, (2) pseudo label prediction, (3) dual guidance scoring, an d (4) selection of highly reliable pseudo labels. In each iteration of the data engine, we employ a geometric accuracy scoring approach to assess pseudo labels based on multiple segmentation metrics, and design a multi-modal agent-driven se mantic scoring approach to enhance the semantic perception of pseudo labels. The se two scoring approaches can effectively improve the reliability of pseudo labe ls by selecting unlabeled samples with higher scores. Our method demonstrates pr omising performance across three mirror detection tasks and exhibits strong gene ralization on unseen examples. Our code will be publicly available at \href http s://github.com/ge-xing/DAM https://github.com/ge-xing/DAM .
****

CH3Depth: Efficient and Flexible Depth Foundation Model with Flow Matching
Jiaqi Li, Yiran Wang, Jinghong Zheng, Junrui Zhang, Liao Shen, Tianqi Liu, Zhigu o Cao; Proceedings of the Computer Vision and Pattern Recognition Conference (CV PR), 2025, pp. 7222-7232
Depth estimation is a fundamental task in 3D vision. An ideal depth estimation m odel is expected to embrace meticulous detail, temporal consistency, and high ef ficiency. Although existing foundation models can perform well in certain specif ic aspects, most of them fall short of fulfilling all the above requirements sim ultaneously. In this paper, we present CH_3Depth, an efficient and flexible mode l for depth estimation with flow matching to address this challenge. Specificall y, 1) we reframe the optimization objective of flow matching as the Inversion by Direct Iteration (InDI) to improve accuracy. 2) To enhance efficiency, we propo se non-uniform sampling to achieve better prediction with fewer sampling steps. 3) We design the Latent Temporal Stabilizer (LTS) to enhance temporal consistenc y by aggregating latent codes of adjacent frames, enabling our method to be ligh tweight and compatible for video depth estimation. CH_3Depth achieves state-of-t he-art performance in zero-shot evaluations across multiple image and video data sets, excelling in prediction accuracy, efficiency, and temporal consistency, hi ghlighting its potential as the next foundation model for depth estimation.
****

Pow3R: Empowering Unconstrained 3D Reconstruction with Camera and Scene Priors
Wonbong Jang, Philippe Weinzaepfel, Vincent Leroy, Lourdes Agapito, Jerome Revau d; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1071-1081
We present Pow3R, a novel large 3D vision regression model that is highly versat ile in the input modalities it accepts. Unlike previous feed-forward models that lack any mechanism to exploit known camera or scene priors at test time, Pow3R incorporates any combination of auxiliary information such as intrinsics, relati ve pose, dense or sparse depth, alongside input images, within a single network. Building upon the recent DUSt3R paradigm, a transformer-based architecture that leverages powerful pre-training, our lightweight and versatile conditioning act s as additional guidance for the network to predict more accurate estimates when auxiliary information is available. During training we feed the model with rand om subsets of modalities at each iteration, which enables the model to operate u nder different levels of known priors at test time. This in turn opens up new ca pabilities, such as performing inference in native image resolution, or point-cl oud completion. Our experiments on 3D reconstruction, depth completion, multi-vi ew depth prediction, multi-view stereo, and multi-view pose estimation tasks yie ld state-of-the-art results and confirm the effectiveness of Pow3R at exploiting all available information. The project webpage is https://europe.naverlabs.com/ pow3r.
****

Efficient Visual State Space Model for Image Deblurring
Lingshun Kong, Jiangxin Dong, Jinhui Tang, Ming-Hsuan Yang, Jinshan Pan; Proceed ings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12710-12719

Convolutional neural networks (CNNs) and Vision Transformers (ViTs) have achieved excellent performance in image restoration. While ViTs generally outperform CNNs by effectively capturing long-range dependencies and input-specific characteristics, their computational complexity increases quadratically with image resolution. This limitation hampers their practical application in high-resolution image restoration. In this paper, we propose a simple yet effective visual state space model (EVSSM) for image deblurring, leveraging the benefits of state space models (SSMs) to visual data. In contrast to existing methods that employ several fixed-direction scanning for feature extraction, which significantly increases the computational cost, we develop an efficient visual scan block that applies various geometric transformations before each SSM-based module, capturing useful non-local information and maintaining high efficiency. In addition, to more effectively capture and represent local information, we propose an efficient discriminative frequency domain-based feedforward network (EDFFN) which can effectively estimate useful frequency information for latent clear image restoration. Extensive experimental results show that the proposed EVSSM performs favorably against state-of-the-art methods on benchmark datasets and real-world images.

****

4D LangSplat: 4D Language Gaussian Splatting via Multimodal Large Language Models

Wanhua Li, Renping Zhou, Jiawei Zhou, Yingwei Song, Johannes Herter, Minghan Qin, Gao Huang, Hanspeter Pfister; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22001-22011

Learning 4D language fields to enable time-sensitive, open-ended language queries in dynamic scenes is essential for many real-world applications. While LangSplat successfully grounds CLIP features into 3D Gaussian representations, achieving precision and efficiency in 3D static scenes, it lacks the ability to handle dynamic 4D fields as CLIP, designed for static image-text tasks, cannot capture temporal dynamics in videos. Real-world environments are inherently dynamic, with object semantics evolving over time. Building a precise 4D language field necessitates obtaining pixel-aligned, object-wise video features, which current vision models struggle to achieve. To address these challenges, we propose 4D LangSplat, which learns 4D language fields to handle time-agnostic or time-sensitive open-vocabulary queries in dynamic scenes efficiently. 4D LangSplat bypasses learning the language field from vision features and instead learns directly from text generated from object-wise video captions via Multimodal Large Language Models (MLLMs). Specifically, we propose a multimodal object-wise video prompting method, consisting of visual and text prompts that guide MLLMs to generate detailed, temporally consistent, high-quality captions for objects throughout a video. These captions are encoded using a Large Language Model into high-quality sentence embeddings, which then serve as pixel-aligned, object-specific feature supervision, facilitating open-vocabulary text queries through shared embedding spaces. Recognizing that objects in 4D scenes exhibit smooth transitions across states, we further propose a status deformable network to model these continuous changes over time effectively. Our results across multiple benchmarks demonstrate that 4D LangSplat attains precise and efficient results for both time-sensitive and time-agnostic open-vocabulary queries.

****

MambaVLT: Time-Evolving Multimodal State Space Model for Vision-Language Tracking

Xinqi Liu, Li Zhou, Zikun Zhou, Jianqiu Chen, Zhenyu He; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8731-8741

The vision-language tracking task aims to perform object tracking based on various modality references. Existing Transformer-based vision-language tracking methods have made remarkable progress by leveraging the global modeling ability of self-attention. However, current approaches still face challenges in effectively exploiting the temporal information and dynamically updating reference features during tracking. Recently, the State Space Model (SSM), known as Mamba, has shown astonishing ability in efficient long-sequence modeling. Particularly, its state space evolving process demonstrates promising capabilities in memorizing mult

imodal temporal information with linear complexity. Witnessing its success, we p
ropose a Mamba-based vision-language tracking model to exploit its state space e
volving ability in temporal space for robust multimodal tracking, dubbed MambaVL
T. In particular, our approach mainly integrates a time-evolving hybrid state sp
ace block and a selective locality enhancement block, to capture contextual info
rmation for multimodal modeling and adaptive reference feature update. Besides,
we introduce a modality-selection module that dynamically adjusts the weighting
between visual and language references, mitigating potential ambiguities from ei
ther reference type. Extensive experimental results show that our method perform
s favorably against state-of-the-art trackers across diverse benchmarks.
****

Enhancing 3D Gaze Estimation in the Wild using Weak Supervision with Gaze Follow
ing Labels
Pierre Vuillecard, Jean-Marc Odobez; Proceedings of the Computer Vision and Patt
ern Recognition Conference (CVPR), 2025, pp. 13508-13518
Accurate 3D gaze estimation in unconstrained real-world environments remains a s
ignificant challenge due to variations in appearance, head pose, occlusion, and
the limited availability of in-the-wild 3D gaze datasets. To address these chall
enges, we introduce a novel Self-Training Weakly-Supervised Gaze Estimation fram
ework (ST-SWGE). This two-stage learning framework leverages diverse 2D gaze dat
asets, such as gaze-following data, which offer rich variations in appearances,
natural scenes, and gaze distributions, and proposes an approach to generate 3D
pseudo-labels and enhance model generalization. Furthermore, traditional modalit
y-specific models, designed separately for images or videos, limit the effective
 use of available training data. To overcome this, we propose the Gaze Transform
er (GaT), a modality-agnostic architecture capable of simultaneously learning st
atic and dynamic gaze information from both image and video datasets. By combini
ng 3D video datasets with 2D gaze target labels from gaze following tasks, our a
pproach achieves the following key contributions: (i) Significant state-of-the-a
rt improvements in within-domain and cross-domain generalization on unconstraine
d benchmarks like Gaze360 and GFIE, with notable cross-modal gains in video gaze
 estimation; (ii) Superior cross-domain performance on datasets such as MPIIFace
Gaze and Gaze360 compared to frontal face methods. Code and pre-trained models w
ill be released to the community.
****

Reward Fine-Tuning Two-Step Diffusion Models via Learning Differentiable Latent-
Space Surrogate Reward
Zhiwei Jia, Yuesong Nan, Huixi Zhao, Gengdai Liu; Proceedings of the Computer Vi
sion and Pattern Recognition Conference (CVPR), 2025, pp. 12912-12922
Recent research has shown that fine-tuning diffusion models (DMs) with arbitrary
 rewards, including non-differentiable ones, is feasible with reinforcement lear
ning (RL) techniques, enabling flexible model alignment. However, applying exist
ing RL methods to step-distilled DMs is challenging for ultra-fast (<=2-step) im
age generation. Our analysis suggests several limitations of policy-based RL met
hods such as PPO or DPO toward this goal. Based on the insights, we propose fine
-tuning DMs with learned differentiable surrogate rewards. Our method, named LaS
RO, learns surrogate reward models in the latent space of SDXL to convert arbitr
ary rewards into differentiable ones for effective reward gradient guidance. LaS
RO leverages pre-trained latent DMs for reward modeling and tailors reward optim
ization for <=2-step image generation with efficient off-policy exploration. LaS
RO is effective and stable for improving ultra-fast image generation with differ
ent reward objectives, outperforming popular RL methods including DDPO and Diffu
sion-DPO. We further show LaSRO's connection to value-based RL, providing theore
tical insights. See our webpage at https://sites.google.com/view/lasro.
****

Detecting Out-of-Distribution Through the Lens of Neural Collapse
Litian Liu, Yao Qin; Proceedings of the Computer Vision and Pattern Recognition
Conference (CVPR), 2025, pp. 15424-15433
Out-of-Distribution (OOD) detection is critical for safe deployment; however, ex
isting detectors often struggle to generalize across datasets of varying scales

and model architectures, and some can incur high computational costs in real-world applications. Inspired by the phenomenon of Neural Collapse, we propose a versatile and efficient OOD detection method. Specifically, we re-characterize prior observations that in-distribution (ID) samples form clusters, demonstrating that, with appropriate centering, these clusters align closely with model weight vectors. Additionally, we reveal that ID features tend to expand into a simplex Equiangular Tight Frame, explaining the common observation that ID features are situated farther from the origin than OOD features. Incorporating both insights from Neural Collapse, our OOD detector leverages feature proximity to weight vectors and complements this approach by using feature norms to effectively filter out OOD samples. Extensive experiments on off-the-shelf models demonstrate the robustness of our OOD detector across diverse scenarios, mitigating generalization discrepancies and enhancing overall performance, with inference latency comparable to that of the basic softmax-confidence detector. Code is available at: https://github.com/litianliu/NCI-OOD.

****

Adaptive Parameter Selection for Tuning Vision-Language Models

Yi Zhang, Yi-Xuan Deng, Meng-Hao Guo, Shi-Min Hu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4280-4290

Vision-language models (VLMs) like CLIP have been widely used in various specific tasks.Parameter-efficient fine-tuning (PEFT) methods, such as prompt and adapter tuning,have become key techniques for adapting these models to specific domains.However, existing approaches rely on prior knowledgeto manually identify the locations requiring fine-tuning.Adaptively selecting which parameters in VLMs should be tuned remains unexplored. In this paper, we propose CLIP with Adaptive Selective Tuning (CLIP-AST), which can be used to automatically select critical parameters in VLMs for fine-tuning for specific tasks.It opportunely leveragesthe adaptive learning rate in the optimizer and improves model performance without extra parameter overhead. We conduct extensive experiments on 13 benchmarks, such as ImageNet, Food101, Flowers102, etc,with different settings, including few-shot learning, base-to-novel class generalization, and out-of-distribution. The results show that CLIP-AST consistently outperforms the original CLIP model as well as its variantsand achieves state-of-the-art (SOTA) performance in all cases. For example, with the 16-shot learning, CLIP-AST surpasses GraphAdapter and PromptSRC by 3.56% and 2.20% in average accuracy on 11 datasets, respectively.Code will be publicly available.

****

MotionMap: Representing Multimodality in Human Pose Forecasting

Reyhaneh Hosseininejad, Megh Shukla, Saeed Saadatnejad, Mathieu Salzmann, Alexandre Alahi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22680-22689

Human pose forecasting is inherently multimodal since multiple future motions exist for an observed pose sequence. However, learning this multimodality is challenging since the task is ill-posed. To address this issue, we propose an alternative paradigm to make the task well-posed. Additionally, while state-of-the-art methods predict multimodality, this is attained through a large volume of predictions obtained by oversampling. However, such an approach glosses over key questions: (1) Can we capture multimodality by efficiently sampling a smaller number of predictions? (2) Subsequently, which of the predicted futures is more likely for an observed pose sequence? We address these questions with MotionMap, a simple yet effective heatmap based representation for multimodality. We extend heatmaps to represent a spatial distribution over the space of all possible motions, where different local maxima correspond to different forecasts for a given observation. Not only can MotionMap capture a variable number of modes per observation, but it also provides confidence measures for different modes. Further, MotionMap captures rare modes that are non-trivial to evaluate yet critical for robustness. Finally, MotionMap allows us to introduce the notion of uncertainty and controllability over the forecasted pose sequence. We support our claims through multiple qualitative and quantitative experiments using popular 3D human pose datasets: Human3.6M and AMASS, highlighting the strengths and limitations of our pr

oposed method.
****
Learning-enabled Polynomial Lyapunov Function Synthesis via High-Accuracy Counterexample-Guided Framework
Hanrui Zhao, Niuniu Qi, Mengxin Ren, Banglong Liu, Shuming Shi, Zhengfeng Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10275-10284
Polynomial Lyapunov function \mathcal V (x) provides mathematically rigorous that converts stability analysis into efficiently solvable optimization problem. Traditional numerical methods rely on user-defined templates, while emerging neural \mathcal V (x) offer flexibility but exhibit poor generalization yield from naive Square polynomial networks. In this paper, we propose a novel learning-enabled polynomial \mathcal V (x) synthesis approach, where a data-driven machine learning process guided by target-based sampling to fit candidate \mathcal V (x) which naturally compatible with the sum-of-squares (SOS) soundness verification. The framework is structured as an iterative loop between a Learner and a Verifier , where the Learner trains expressive polynomial \mathcal V (x) network via polynomial expansions, while the Verifier encodes learned candidates with SOS constraints to identify a real \mathcal V (x) by solving LMI feasibility test problems. The entire procedure is driven by a high-accuracy counterexample guidance technique to further enhance efficiency. Experimental results demonstrate that our approach outperforms both SMT-based polynomial neural Lyapunov function synthesis and traditional SOS method.
****
Factored-NeuS: Reconstructing Surfaces, Illumination, and Materials of Possibly Glossy Objects
Yue Fan, Ningjing Fan, Ivan Skorokhodov, Oleg Voynov, Savva Ignatyev, Evgeny Burnaev, Peter Wonka, Yiqun Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21317-21327
We develop a method that recovers the surface, materials, and illumination of a scene from its posed multi-view images. In contrast to prior work, it does not require any additional data and can handle glossy objects or bright lighting. It is a progressive inverse rendering approach, which consists of three stages. In the first stage, we reconstruct the scene radiance and signed distance function (SDF) with a novel regularization strategy for specular reflections. We propose to explain a pixel color using both surface and volume rendering jointly, which allows for handling complex view-dependent lighting effects for surface reconstruction. In the second stage, we distill light visibility and indirect illumination from the learned SDF and radiance field using learnable mapping functions. Finally, we design a method for estimating the ratio of incoming direct light reflected in a specular manner and use it to reconstruct the materials and direct illumination. Experimental results demonstrate that the proposed method outperforms the current state-of-the-art in recovering surfaces, materials, and lighting without relying on any additional data.
****
GaussianSpa: An "Optimizing-Sparsifying" Simplification Framework for Compact and High-Quality 3D Gaussian Splatting
Yangming Zhang, Wenqi Jia, Wei Niu, Miao Yin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26673-26682
3D Gaussian Splatting (3DGS) has emerged as a mainstream for novel view synthesis, leveraging continuous aggregations of Gaussian functions to model scene geometry. However, 3DGS suffers from substantial memory requirements to store the large amount of Gaussians, hindering its efficiency and practicality. To address this challenge, we introduce GaussianSpa, an optimization-based simplification framework for compact and high-quality 3DGS. Specifically, we formulate the simplification objective as a constrained optimization problem associated with the 3DGS training. Correspondingly, we propose an efficient "optimizing-sparsifying" solution for the formulated problem, alternately solving two independent sub-problems and gradually imposing substantial sparsity onto the Gaussians in the 3DGS training process. We conduct quantitative and qualitative evaluations on various d

atasets, demonstrating the superiority of GaussianSpa over existing state-of-the
-art approaches. Notably, GaussianSpa achieves an average PSNR improvement of 0.
9 dB on the real-world Deep Blending dataset with 10X fewer Gaussians compared t
o the vanilla 3DGS.
****

Sparse2DGS: Geometry-Prioritized Gaussian Splatting for Surface Reconstruction f
rom Sparse Views
Jiang Wu, Rui Li, Yu Zhu, Rong Guo, Jinqiu Sun, Yanning Zhang; Proceedings of th
e Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11307-113
16
We present a Gaussian Splatting method for surface reconstruction using sparse i
nput views. Previous methods relying on dense views struggle with extremely spar
se Structure-from-Motion points for initialization. While learning-based Multi-v
iew Stereo (MVS) provides dense 3D points, directly combining it with Gaussian S
platting leads to suboptimal results due to the ill-posed nature of sparse-view
geometric optimization. We propose Sparse2DGS, an MVS-initialized Gaussian Splat
ting pipeline for complete and accurate reconstruction. Our key insight is to in
corporate the geometric-prioritized enhancement schemes, allowing for direct and
 robust geometric learning under ill-posed conditions. Sparse2DGS outperforms ex
isting methods by notable margins, with 1.13 Chamfer Distance error compared to
2DGS (2.81) on the DTU dataset using 3 views. Meanwhile, our method is 2x faster
 than NeRF-based fine-tuning approach. Code is available at https://github.com/W
uuu3511/Sparse2DGS.
****

VinTAGe: Joint Video and Text Conditioning for Holistic Audio Generation
Saksham Singh Kushwaha, Yapeng Tian; Proceedings of the Computer Vision and Patt
ern Recognition Conference (CVPR), 2025, pp. 13529-13539
Recent advances in audio generation have focused on text-to-audio (T2A) and vide
o-to-audio (V2A) tasks. However, T2A or V2A methods cannot generate holistic sou
nds (onscreen and off-screen). This is because T2A cannot generate sounds aligni
ng with onscreen objects, while V2A cannot generate semantically complete (offsc
reen sounds missing). In this work, we address the task of holistic audio genera
tion: given a video and a text prompt, we aim to generate both onscreen and offs
creen sounds that are temporally synchronized with the video and semantically al
igned with text and video. Previous approaches for joint text and video-to-audio
 generation often suffer from modality bias, favoring one modality over the othe
r. To overcome this limitation, we introduce VinTAGe, a flow-based transformer m
odel that jointly considers text and video to guide audio generation. Our framew
ork comprises two key components: a Visual-Text Encoder and a Joint VT-SiT model
. To reduce modality bias and improve generation quality, we employ pretrained u
ni-modal text-to-audio and video-to-audio generation models for additional guida
nce. Due to the lack of appropriate benchmarks, we also introduce VinTAGe-Bench,
 a dataset of 636 video-text-audio pairs containing both onscreen and offscreen
sounds. Our comprehensive experiments on VinTAGe-Bench demonstrate that joint te
xt and visual interaction is necessary for holistic audio generation. Furthermor
e, VinTAGe achieves state-of-the-art results on the VGGSound benchmark.We will r
elease our pretrained models and the VinTAGe-Bench dataset to facilitate future
research in this exciting field.
****

Efficient Decoupled Feature 3D Gaussian Splatting via Hierarchical Compression
Zhenqi Dai, Ting Liu, Yanning Zhang; Proceedings of the Computer Vision and Patt
ern Recognition Conference (CVPR), 2025, pp. 11156-11166
Efficient 3D scene representation has become a key challenge with the rise of 3D
 Gaussian Splatting (3DGS), particularly when incorporating semantic information
 into the scene representation. Existing 3DGS-based methods embed both color and
 high-dimensional semantic features into a single field, leading to significant
 storage and computational overhead. To mitigate this, we propose Decoupled Feat
ure 3D Gaussian Splatting (DF-3DGS), a novel method that decouples the color and
 semantic fields, thereby reducing the number of 3D Gaussians required for seman
tic representation. We then introduce a hierarchical compression strategy that f

irst employs our novel quantization approach with dynamic codebook evolution to reduce data size, followed by a scene-specific autoencoder for further compression of the semantic feature dimensions. This multi-stage approach results in a compact representation that enhances both storage efficiency and reconstruction speed. Experimental results demonstrate that DF-3DGS outperforms previous 3DGS-based methods, achieving faster training and rendering times while requiring less storage, without sacrificing performance--in fact, it improves performance in the novel view semantic segmentation task. Specifically, DF-3DGS achieves remarkable improvements over Feature 3DGS, reducing training time by 10xand storage by 20x, while improving the mIoU of novel view semantic segmentation by 4%. The code will be publicly available.

****

CountLLM: Towards Generalizable Repetitive Action Counting via Large Language Model

Ziyu Yao, Xuxin Cheng, Zhiqi Huang, Lei Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19143-19153

Repetitive action counting, which aims to count periodic movements in a video, is valuable for video analysis applications such as fitness monitoring. However, existing methods largely rely on regression networks with limited representational capacity, which hampers their ability to accurately capture variable periodic patterns. Additionally, their supervised learning on narrow, limited training sets leads to overfitting and restricts their ability to generalize across diverse scenarios. To address these challenges, we propose CountLLM, the first large language model (LLM)-based framework that takes video data and periodic text prompts as inputs and outputs the desired counting value. CountLLM leverages the rich clues from explicit textual instructions and the powerful representational capabilities of pre-trained LLMs for repetitive action counting. To effectively guide CountLLM, we develop a periodicity-based structured template for instructions that describes the properties of periodicity and implements a standardized answer format to ensure consistency. Additionally, we propose a progressive multimodal training paradigm to enhance the periodicity-awareness of the LLM. Empirical evaluations on widely recognized benchmarks demonstrate CountLLM's superior performance and generalization, particularly in handling novel and out-of-domain actions that deviate significantly from the training data, offering a promising avenue for repetitive action counting.

****

Navigating the Unseen: Zero-shot Scene Graph Generation via Capsule-Based Equivariant Features

Wenhuan Huang, Yi JI, Guiqian Zhu, Li Ying, Chunping Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29448-29457

In scene graph generation (SGG), the accurate prediction of unseen triples is essential for its effectiveness in downstream vision-language tasks. We hypothesize that the predicates of unseen triples can be viewed as transformations of seen predicates in feature space, and the essence of the zero-shot task is to bridge the gap caused by this transformation. Traditional models, however, have difficulty addressing this challenge, which we attribute to their inability to model the predicates equivariant. To overcome this limitation, we introduce a novel framework based on capsule networks (CAPSGG). We propose a Three-Stream Pipeline that generates modality-specific representations for predicates, while building low-level predicate capsules of these modalities. Then these capsules are aggregated into high-level predicate capsules using a Routing Capsule Layer. In addition, we introduce GroupLoss to aggregate capsules with the same predicate label into groups. This replaces the global loss with the intra-group loss, effectively balancing the learning of predicate invariance and equivariant features, while mitigating the impact of the severe long-tail distribution of the predicate categories. Our extensive experiments demonstrate the notable superiority of our approach over state-of-the-art methods, with zero-shot indicators outperforming up to 132.26% on SGCls task than the T-CAR [21]. Our code will be available upon publication.

****

VL-RewardBench: A Challenging Benchmark for Vision-Language Generative Reward Models
Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, Lingpeng Kong, Qi Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24657-24668
Vision-language generative reward models (VL-GenRMs) play a crucial role in aligning and evaluating multimodal AI systems, yet their own evaluation remains under-explored. Current assessment methods primarily rely on AI-annotated preference labels from traditional VL tasks, which can introduce biases and often fail to effectively challenge state-of-the-art models. To address these limitations, we introduce VL-RewardBench, a comprehensive benchmark spanning general multimodal queries, visual hallucination detection, and complex reasoning tasks. Through our AI-assisted annotation pipeline that combines sample selection with human verification, we curate 1,250 high-quality examples specifically designed to probe VL-GenRMs limitations. Comprehensive evaluation across 16 leading large vision-language models demonstrates VL-RewardBench's effectiveness as a challenging testbed, where even GPT-4o achieves only 65.4% accuracy, and state-of-the-art open-source models such as Qwen2-VL-72B, struggle to surpass random-guessing. Importantly, performance on VL-RewardBench strongly correlates (Pearson's r > 0.9) with MMMU-Pro accuracy using Best-of-N sampling with VL-GenRMs. Analysis experiments uncover three critical insights for improving VL-GenRMs: (i) models predominantly fail at basic visual perception tasks rather than reasoning tasks; (ii) inference-time scaling benefits vary dramatically by model capacity; and (iii) training VL-GenRMs to learn to judge substantially boosts judgment capability (+14.7% accuracy for a 7B VL-GenRM). We believe VL-RewardBench along with the experimental insights will become a valuable resource for advancing VL-GenRMs. Project page: https://vl-rewardbench.github.io
****
ASHiTA: Automatic Scene-grounded HIerarchical Task Analysis
Yun Chang, Leonor Fermoselle, Duy Ta, Bernadette Bucher, Luca Carlone, Jiuguang Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29458-29468
While recent work in scene reconstruction and understanding has made strides in grounding natural language to physical 3D environments, it is still challenging to ground abstract, high-level instructions to a 3D scene. High-level instructions might not explicitly invoke semantic elements in the scene, and even the process of breaking a high-level task into a set of more concrete subtasks --a process called hierarchical task analysis-- is environment-dependent. In this work, we propose ASHiTA, the first framework that generates a task hierarchy grounded to a 3D scene graph by breaking down high-level tasks into grounded subtasks. ASHiTA alternates LLM-assisted hierarchical task analysis --to generate the task breakdown-- with task-driven scene graph construction to generate a suitable representation of the environment. Our experiments show that ASHiTA performs significantly better than LLM baselines in breaking down high-level tasks into environment-dependent subtasks and is additionally able to achieve grounding performance comparable to state-of-the-art methods
****
Patient-Level Anatomy Meets Scanning-Level Physics: Personalized Federated Low-Dose CT Denoising Empowered by Large Language Model
Ziyuan Yang, Yingyu Chen, Zhiwen Wang, Hongming Shan, Yang Chen, Yi Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5154-5163
Reducing radiation doses benefits patients, but the resultant low-dose computed tomography (LDCT) images often suffer from clinically unacceptable noise and artifacts. While deep learning (DL) has shown promise in LDCT reconstruction, it requires large-scale data collection from multiple clients, raising privacy concerns. Federated learning (FL) has been introduced to mitigate these privacy concerns; however, current methods are typically tailored to specific scanning protocols, which limits their generalizability and makes them less effective for unseen

protocols. To address these issues, we propose SCAN-PhysFed, a novel SCanning- and ANatomy-level personalized Physics-Driven Federated learning paradigm for LD CT reconstruction. Since the noise distribution in LDCT data is closely tied to scanning protocols and anatomical structures, we propose a dual-level physics-in formed way to address these challenges. Specifically, we incorporate physical an d anatomical prompts into our physics-informed hypernetworks to capture scanning - and anatomy-specific information, enabling dual-level physics-driven personali zation of imaging features. These prompts are derived from the scanning protocol and the radiology report generated by a medical large language model (MLLM). Su bsequently, client-specific decoders project these dual-level personalized imagi ng features back into the image domain. Besides, to tackle the challenge of unse en data, we introduce a novel protocol vector-quantization strategy (PVQS), whic h ensures consistent performance across new clients by quantifying unseen scanni ng codes to the closest match in the scanning codebook. Extensive experimental r esults demonstrate the superior performance of SCAN-PhysFed on public datasets.
****

Exploiting Deblurring Networks for Radiance Fields
Haeyun Choi, Heemin Yang, Janghyeok Han, Sunghyun Cho; Proceedings of the Comput er Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6012-6021
In this paper, we propose DeepDeblurRF, a novel radiance field deblurring approa ch that can synthesize high-quality novel views from blurred training views with significantly reduced training time. DeepDeblurRF leverages deep neural network (DNN)-based deblurring modules to enjoy their deblurring performance and comput ational efficiency. To effectively combine DNN-based deblurring and radiance fie ld construction, we propose a novel radiance field (RF)-guided deblurring and an iterative framework that performs RF-guided deblurring and radiance field const ruction in an alternating manner. Moreover, DeepDeblurRF is compatible with vari ous scene representations, such as voxel grids and 3D Gaussians, expanding its a pplicability. We also present BlurRF-Synth, the first large-scale synthetic data set for training radiance field deblurring frameworks. We conduct extensive expe riments on both camera motion blur and defocus blur, demonstrating that DeepDebl urRF achieves state-of-the-art novel-view synthesis quality with significantly r educed training time.
****

Rethinking Lanes and Points in Complex Scenarios for Monocular 3D Lane Detection
Yifan Chang, Junjie Huang, Xiaofeng Wang, Yun Ye, Zhujin Liang, Yi Shan, Dalong Du, Xingang Wang; Proceedings of the Computer Vision and Pattern Recognition Con ference (CVPR), 2025, pp. 6802-6811
Monocular 3D lane detection is a fundamental task in autonomous driving. Althoug h sparse-point methods lower computational load and maintain high accuracy in co mplex lane geometries, current methods fail to fully leverage the geometric stru cture of lanes in both lane geometry representations and model design. In lane g eometry representations, we present a theoretical analysis alongside experimenta l validation to verify that current sparse lane representation methods contain i nherent flaws, resulting in potential errors of up to 20 m, which raise signific ant safety concerns for driving. To address this issue, we propose a novel patch ing strategy to completely represent the full lane structure. To enable existing models to match this strategy, we introduce the EndPoint head (EP-head), which adds a patching distance to endpoints. The EP-head enables the model to predict more complete lane representations even with fewer preset points, effectively ad dressing existing limitations and paving the way for models that are faster and require fewer parameters in the future. In model design, to enhance the model's perception of lane structures, we propose the PointLane attention (PL-attention) , which incorporates prior geometric knowledge into the attention mechanism. Ext ensive experiments demonstrate the effectiveness of the proposed methods on vari ous state-of-the-art models. For instance, in terms of the overall F1-score, our methods improve Persformer by 4.4 points, Anchor3DLane by 3.2 points, and LATR by 2.8 points. The code will be available soon.
****

SPAR3D: Stable Point-Aware Reconstruction of 3D Objects from Single Images

Zixuan Huang, Mark Boss, Aaryaman Vasishta, James M. Rehg, Varun Jampani; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16860-16870

We study the problem of single-image 3D object reconstruction. Recent works have diverged into two directions: regression-based modeling and generative modeling. Regression methods efficiently infer visible surfaces, but struggle with occluded regions. Generative methods handle uncertain regions better by modeling distributions, but are computationally expensive and the generation is often misaligned with visible surfaces. In this paper, we present SPAR3D, a novel two-stage approach aiming to take the best of both directions. The first stage of SPAR3D generates sparse 3D point clouds using a lightweight point diffusion model, which has a fast sampling speed. The second stage uses both the sampled point cloud and the input image to create highly detailed meshes. Our two-stage design enables probabilistic modeling of the ill-posed single-image 3D task while maintaining high computational efficiency and great output fidelity. Using point clouds as an intermediate representation further allows for interactive user edits. Evaluated on diverse datasets, SPAR3D demonstrates superior performance over previous state-of-the-art methods, at an inference speed of 0.7 seconds. Project page with code and model: https://spar3d.github.io

****

Discovering Fine-Grained Visual-Concept Relations by Disentangled Optimal Transport Concept Bottleneck Models

Yan Xie, Zequn Zeng, Hao Zhang, Yucheng Ding, Yi Wang, Zhengjue Wang, Bo Chen, Hongwei Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30199-30209

Concept Bottleneck Models (CBMs) try to make the decision-making process transparent by exploring an intermediate concept space between the input image and the output prediction. Existing CBMs just learn coarse-grained relations between the whole image and the concepts, less considering local image information, leading to two main drawbacks: i) they often produce spurious visual-concept relations, hence decreasing model reliability; and ii) though CBMs could explain the importance of every concept to the final prediction, it is still challenging to tell which visual region produces the prediction. To solve these problems, this paper proposes a Disentangled Optimal Transport CBM (DOT-CBM) framework to explore fine-grained visual-concept relations between local image patches and concepts. Specifically, we model the concept prediction process as a transportation problem between the patches and concepts, thereby achieving explicit fine-grained feature alignment. We also incorporate orthogonal projection losses within the modality to enhance local feature disentanglement. To further address the shortcut issues caused by statistical biases in the data, we utilize the visual saliency map and concept label statistics as transportation priors. Thus, DOT-CBM can visualize inversion heatmaps, provide more reliable concept predictions, and produce more accurate class predictions. Comprehensive experiments demonstrate that our proposed DOT-CBM achieves SOTA performance on several tasks, including image classification, local part detection and out-of-distribution generalization.

****

Focus-N-Fix: Region-Aware Fine-Tuning for Text-to-Image Generation

Xiaoying Xing, Avinab Saha, Junfeng He, Susan Hao, Paul Vicol, Moonkyung Ryu, Gang Li, Sahil Singla, Sarah Young, Yinxiao Li, Feng Yang, Deepak Ramachandran; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18486-18496

Text-to-image (T2I) generation has made significant advances in recent years, but challenges still remain in the generation of perceptual artifacts, misalignment with complex prompts, and safety. The prevailing approach to address these issues involves collecting human feedback on generated images, training reward models to estimate human feedback, and then fine-tuning T2I models based on the reward models to align them with human preferences. However, while existing reward fine-tuning methods can produce images with higher rewards, they may change model behavior in unexpected ways. For example, fine-tuning for one quality aspect (e.g., safety) may degrade other aspects (e.g., prompt alignment), or may lead to

reward hacking (e.g., finding a way to increase rewards without having the inten
ded effect). In this paper, we propose Focus-N-Fix, the first region-aware fine-
tuning method that trains models to correct only previously problematic image re
gions. The resulting fine-tuned model generates images with the same high-level
structure as the original model but shows significant improvements in regions wh
ere the original model was deficient in safety (over-sexualization and violence)
, plausibility, or other criteria. Our experiments demonstrate that Focus-N-Fix
improves these localized quality aspects with little or no degradation to others
 and typically imperceptible changes in the rest of the image. Disclaimer: This
paper contains images that may be overly sexual, violent, offensive or harmful.
****

RoomTour3D: Geometry-Aware Video-Instruction Tuning for Embodied Navigation
Mingfei Han, Liang Ma, Kamila Zhumakhanova, Ekaterina Radionova, Jingyi Zhang, X
iaojun Chang, Xiaodan Liang, Ivan Laptev; Proceedings of the Computer Vision and
 Pattern Recognition Conference (CVPR), 2025, pp. 27586-27596
Vision-and-Language Navigation (VLN) suffers from the limited diversity and scal
e of training data, primarily constrained by the manual curation of existing sim
ulators.To address this, we introduce RoomTour3D, a video-instruction dataset de
rived from web-based room tour videos that capture real-world indoor spaces and
human walking demonstrations. Unlike existing VLN datasets, RoomTour3D leverages
 the scale and diversity of online videos to generate open-ended human walking t
rajectories and open-world navigable instructions. To compensate for the lack of
 navigation data in online videos, we perform 3D reconstruction and obtain 3D tr
ajectories of walking paths augmented with additional information on the room ty
pes, object locations and 3D shape of surrounding scenes. Our dataset includes ~
100K open-ended description-enriched trajectories with ~200K instructions, and 1
7K action-enriched trajectories from 1847 room tour environments.We demonstrate
experimentally that RoomTour3D enables significant improvements across multiple
VLN tasks including CVDN, SOON, R2R, and REVERIE.Moreover, RoomTour3D facilitate
s the development of trainable zero-shot VLN agents, showcasing the potential an
d challenges of advancing towards open-world navigation.
****

PSA-SSL: Pose and Size-aware Self-Supervised Learning on LiDAR Point Clouds
Barza Nisar, Steven L. Waslander; Proceedings of the Computer Vision and Pattern
 Recognition Conference (CVPR), 2025, pp. 6670-6679
Self-supervised learning (SSL) on 3D point clouds has the potential to learn fea
ture representations that can transfer to diverse sensors and multiple downstrea
m perception tasks. However, recent SSL approaches fail to define pretext tasks
that retain geometric information such as object pose and scale, which can be d
etrimental to the performance of downstream localization and geometry-sensitive
3D scene understanding tasks, such as 3D semantic segmentation and 3D object det
ection. We propose PSA-SSL, a novel extension to point cloud SSL that learns obj
ect pose and size-aware (PSA) features. Our approach defines a self-supervised b
ounding box regression pretext task, which retains object pose and size informat
ion. Furthermore, we incorporate LiDAR beam pattern augmentation on input point
clouds, which encourages learning sensor-agnostic features. Our experiments demo
nstrate that with a single pretrained model, our light-weight yet effective exte
nsions achieve significant improvements on 3D semantic segmentation with limited
 labels across popular autonomous driving datasets (Waymo, nuScenes, SemanticKIT
TI). Moreover, our approach outperforms other state-of-the-art SSL methods on 3D
 semantic segmentation (using up to 10 times less labels), as well as on 3D obje
ct detection. Our code will be released on https://github.com/TRAILab/PSA-SSL.
****

Bringing CLIP to the Clinic: Dynamic Soft Labels and Negation-Aware Learning for
 Medical Analysis
Hanbin Ko, Chang-Min Park; Proceedings of the Computer Vision and Pattern Recogn
ition Conference (CVPR), 2025, pp. 25897-25906
The development of large-scale image-text pair datasets has significantly advanc
ed self-supervised learning in Vision-Language Processing (VLP). However, direct
ly applying general-domain architectures such as CLIP to medical data presents c

hallenges, particularly in handling negations and addressing the inherent data i
mbalance of medical datasets. To address these issues, we propose a novel approa
ch that integrates clinically-enhanced dynamic soft labels and medical graphical
 alignment, thereby improving clinical comprehension and improving the applicabi
lity of contrastive loss in medical contexts. Furthermore, we introduce negation
-based hard negatives to deepen the model's understanding of the complexities of
 clinical language. Our approach is easily integrated into medical CLIP training
 pipeline and achieves state-of-the-art performance across multiple tasks, inclu
ding zero-shot, fine-tuned classification and report retrieval. To comprehensive
ly evaluate our model's capacity in understanding clinical language, we introduc
e CXR-Align, a benchmark uniquely designed to evaluate the understanding of nega
tion and clinical information within chest X-ray (CXR) datasets. Experimental re
sults demonstrate that our proposed methods are straightforward to implement and
 generalize effectively across contrastive learning frameworks, enhancing medica
l VLP capabilities and advancing clinical language understanding in medical imag
ing.
****

SnapGen: Taming High-Resolution Text-to-Image Models for Mobile Devices with Eff
icient Architectures and Training

Jierun Chen, Dongting Hu, Xijie Huang, Huseyin Coskun, Arpit Sahni, Aarush Gupta
, Anujraaj Goyal, Dishani Lahiri, Rajesh Singh, Yerlan Idelbayev, Junli Cao, Yan
yu Li, Kwang-Ting Cheng, S.-H. Gary Chan, Mingming Gong, Sergey Tulyakov, Anil K
ag, Yanwu Xu, Jian Ren; Proceedings of the Computer Vision and Pattern Recogniti
on Conference (CVPR), 2025, pp. 7997-8008

Existing text-to-image (T2I) diffusion models face several limitations, includin
g large model sizes, slow runtime, and low-quality generation on mobile devices.
 This paper aims to address all of these challenges by developing an extremely s
mall and fast T2I model that generates high-resolution and high-quality images o
n mobile platforms. We propose several techniques to achieve this goal. First, w
e systematically examine the design choices of the network architecture to reduc
e model parameters and latency, while ensuring high-quality generation. Second,
to further improve generation quality, we employ cross-architecture knowledge di
stillation from a much larger model, using a multi-level approach to guide the t
raining of our model from scratch. Third, we enable a few-step generation by int
egrating adversarial guidance with knowledge distillation. Our model, for the fi
rst time, demonstrates the generation of 1024x1024 px images on a mobile device
in 1.2 to 2.3 seconds. On ImageNet-1K, our model, with only 372M parameters, ach
ieves an FID of 2.06 for 256x256 px generation. On T2I benchmarks (i.e., GenEval
 and DPG-Bench), our model with merely 379M parameters surpasses large-scale mod
els with billions of parameters at a significantly smaller size (e.g., 7x smalle
r than SDXL, 14x smaller than IF-XL).
****

Label Shift Meets Online Learning: Ensuring Consistent Adaptation with Universal
 Dynamic Regret

Yucong Dai, Shilin Gu, Ruidong Fan, Chao Xu, Chenping Hou; Proceedings of the Co
mputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15392-15401

Label shift, which investigates the adaptation of label distributions between th
e fixed source and target domains, has attracted significant research interests
and broad applications in offline settings. In real-world scenarios, however, da
ta often arrives as a continuous stream. Addressing label shift in online learni
ng settings is paramount. Existing strategies, which tailor traditional offline
label shift techniques to online settings, have degraded performance due to the
inconsistent estimation of label distributions and violation of convex assumptio
n for theoretical guarantee. In this paper, we propose a novel method to ensure
consistent adaptation to online label shift. We construct a new convex risk esti
mator that is pivotal for both online optimization and theoretical analysis. Fur
thermore, we enhance an optimistic online algorithm as the base learner and refi
ne the classifier using an ensemble method. Theoretically, we derive a universal
 dynamic regret which achieves minimax optimal. Extensive experiments on both re
al-world datasets and human motion task demonstrate the superiority of our metho

d comparing existing methods.
****

A Physics-Informed Blur Learning Framework for Imaging Systems
Liqun Chen, Yuxuan Li, Jun Dai, Jinwei Gu, Tianfan Xue; Proceedings of the Compu ter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10913-10922
Accurate blur estimation is essential for high-performance imaging across variou s applications. Blur is typically represented by the point spread function (PSF) . In this paper, we propose a physics-informed PSF learning framework for imagin g systems, consisting of a simple calibration followed by a learning process. Ou r framework could achieve both high accuracy and universal applicability. Inspir ed by the Seidel PSF model for representing spatially varying PSF, we identify i ts limitations in optimization and introduce a novel wavefront-based PSF model a ccompanied by an optimization strategy, both reducing optimization complexity an d improving estimation accuracy. Moreover, our wavefront-based PSF model is inde pendent of lens parameters, eliminate the need for prior knowledge of the lens. To validate our approach, we compare it with recent PSF estimation methods (Degr adation Transfer and Fast Two-step) through a deblurring task, where all the est imated PSFs are used to train state-of-the-art deblurring algorithms. Our approa ch demonstrates improvements in image quality in simulation and also showcases n oticeable visual quality improvements on real captured images.
****

A Semantic Knowledge Complementarity based Decoupling Framework for Semi-supervi sed Class-imbalanced Medical Image Segmentation
Zheng Zhang, Guanchun Yin, Bo Zhang, Wu Liu, Xiuzhuang Zhou, Wendong Wang; Proce edings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, p p. 25940-25949
The limited data annotations have made semi-supervised learning (SSL) increasing ly popular in medical image analysis. However, the use of pseudo labels in SSL d egrades the performance of decoders that heavily rely on high-accuracy annotatio ns. This issue is particularly pronounced in class-imbalanced multi-organ segmen tation tasks, where small organs may be under-segmented or even ignored. In this paper, we propose SKCDF, a semantic knowledge complementarity based decoupling framework for multi-organ segmentation in class-imbalanced medical images. SKCDF decouples the data flow based on the responsibilities of the encoder and decode r during model training to make the model effectively learn semantic features, w hile mitigating the negative impact of unlabeled data on the semantic segmentati on task. We also design a semantic knowledge complementarity module that adopts labeled data to guide the generation of pseudo labels and enriches the semantic features of labeled data with unlabeled data, which improves the quality of gene rated pseudo labels and the robustness of the overall model. Furthermore, we des ign an auxiliary balanced segmentation head based training strategy to further e nhance the segmentation performance of small organs. Experimental results on the Synapse and AMOS datasets show that our method significantly outperforms existi ng methods.
****

Community Forensics: Using Thousands of Generators to Train Fake Image Detectors
Jeongsoo Park, Andrew Owens; Proceedings of the Computer Vision and Pattern Reco gnition Conference (CVPR), 2025, pp. 8245-8257
One of the key challenges of detecting AI-generated images is spotting images th at have been created by previously unseen generative models. We argue that the l imited diversity of the training data is a major obstacle to addressing this pro blem, and we propose a new dataset that is significantly larger and more diverse than prior works. As part of creating this dataset, we systematically download thousands of text-to-image latent diffusion models and sample images from them. We also collect images from dozens of popular open source and commercial models. The resulting dataset contains 2.7M images that have been sampled from 4803 dif ferent models. These images collectively capture a wide range of scene content, generator architectures, and image processing settings. Using this dataset, we s tudy the generalization abilities of fake image detectors. Our experiments sugge st that detection performance improves as the number of models in the training s

et increases, even when these models have similar architectures. We also find th
at increasing the diversity of the models improves detection performance, and th
at our trained detectors generalize better than those trained on other datasets.
 The dataset can be found in https://jespark.net/projects/2024/community_forensi
cs
****

ModeSeq: Taming Sparse Multimodal Motion Prediction with Sequential Mode Modelin
g
Zikang Zhou, Hengjian Zhou, Haibo Hu, Zihao Wen, Jianping Wang, Yung-Hui Li, Yu-
Kai Huang; Proceedings of the Computer Vision and Pattern Recognition Conference
 (CVPR), 2025, pp. 1612-1621
Anticipating the multimodality of future events lays the foundation for safe aut
onomous driving. However, multimodal motion prediction for traffic agents has be
en clouded by the lack of multimodal ground truth. Existing works predominantly
adopt the winner-take-all training strategy to tackle this challenge, yet still
suffer from limited trajectory diversity and uncalibrated mode confidence. While
 some approaches address these limitations by generating excessive trajectory ca
ndidates, they necessitate a post-processing stage to identify the most represen
tative modes, a process lacking universal principles and compromising trajectory
 accuracy. We are thus motivated to introduce ModeSeq, a new multimodal predicti
on paradigm that models modes as sequences. Unlike the common practice of decodi
ng multiple plausible trajectories in one shot, ModeSeq requires motion decoders
 to infer the next mode step by step, thereby more explicitly capturing the corr
elation between modes and significantly enhancing the ability to reason about mu
ltimodality. Leveraging the inductive bias of sequential mode prediction, we als
o propose the Early-Match-Take-All (EMTA) training strategy to diversify the tra
jectories further. Without relying on dense mode prediction or heuristic post-pr
ocessing, ModeSeq considerably improves the diversity of multimodal output while
 attaining satisfactory trajectory accuracy, resulting in balanced performance o
n motion prediction benchmarks. Moreover, ModeSeq naturally emerges with the cap
ability of mode extrapolation, which supports forecasting more behavior modes wh
en the future is highly uncertain.
****

Quaffure: Real-Time Quasi-Static Neural Hair Simulation
Tuur Stuyck, Gene Wei-Chin Lin, Egor Larionov, Hsiao-yu Chen, Aljaz Bozic, Nikol
aos Sarafianos, Doug Roble; Proceedings of the Computer Vision and Pattern Recog
nition Conference (CVPR), 2025, pp. 239-249
Realistic hair motion is crucial for high-quality avatars, but it is often limit
ed by the computational resources available for real-time applications. To addre
ss this challenge, we propose a novel neural approach to predict physically plau
sible hair deformations that generalizes to various body poses, shapes, and hair
 styles. Our model is trained using a self-supervised loss, eliminating the need
 for expensive data generation and storage. We demonstrate our method's effectiv
eness through numerous results across a wide range of pose and shape variations,
 showcasing its robust generalization capabilities and temporally smooth results
. Our approach is highly suitable for real-time applications with an inference t
ime of only a few milliseconds on consumer hardware and its ability to scale to
predicting 1000 grooms in 0.3 seconds. Code will be released.
****

Towards Practical Real-Time Neural Video Compression
Zhaoyang Jia, Bin Li, Jiahao Li, Wenxuan Xie, Linfeng Qi, Houqiang Li, Yan Lu; P
roceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 202
5, pp. 12543-12552
We introduce a practical real-time neural video codec (NVC) designed to deliver
high compression ratio, low latency and broad versatility. In practice, the codi
ng speed of NVCs depends on 1) computational costs, and 2) non-computational ope
rational costs, such as memory I/O and the number of function calls. While most
efficient NVCs prioritize reducing computational cost, we identify operational c
ost as the primary bottleneck to achieving higher coding speed. Leveraging this
insight, we introduce a set of efficiency-driven design improvements focused on

minimizing operational costs. Specifically, we employ implicit temporal modeling to eliminate complex explicit motion modules, and use single low-resolution latent representations rather than progressive downsampling. These innovations significantly accelerate NVC without sacrificing compression quality. Additionally, we implement model integerization for consistent cross-device coding and a module-bank-based rate control scheme to improve practical adaptability. Experiments show our proposed DCVC-RT achieves an impressive average encoding/decoding speed at 125.2/112.8 fps (frames per second) for 1080p video, while saving an average of 21% in bitrate compared to H.266/VTM. The code is available at https://github.com/microsoft/DCVC.

****

DepthSplat: Connecting Gaussian Splatting and Depth

Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, Marc Pollefeys; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16453-16463

Gaussian splatting and single-view depth estimation are typically studied in isolation. In this paper, we present DepthSplat to connect Gaussian splatting and depth estimation and study their interactions. More specifically, we first contribute a robust multi-view depth model by leveraging pre-trained monocular depth features, leading to high-quality feed-forward 3D Gaussian splatting reconstructions. We also show that Gaussian splatting can serve as an unsupervised pre-training objective for learning powerful depth models from large-scale multi-view posed datasets. We validate the synergy between Gaussian splatting and depth estimation through extensive ablation and cross-task transfer experiments. Our DepthSplat achieves state-of-the-art performance on ScanNet, RealEstate10K and DL3DV datasets in terms of both depth estimation and novel view synthesis, demonstrating the mutual benefits of connecting both tasks. In addition, DepthSplat enables feed-forward reconstruction from 12 input views (512x960 resolutions) in 0.6 seconds.

****

LumiNet: Latent Intrinsics Meets Diffusion Models for Indoor Scene Relighting

Xiaoyan Xing, Konrad Groh, Sezer Karaoglu, Theo Gevers, Anand Bhattad; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 442-452

We introduce LumiNet, a novel architecture that leverages generative models and latent intrinsic representations for transferring lighting from one image to another. Given a source image and a target lighting image, LumiNet generates a relit version of the source scene that captures the target's lighting. Our approach makes two key contributions: a data curation strategy from the StyleGAN-based relighting model for our training, and a modified diffusion-based ControlNet that processes both latent intrinsic properties from the source image and latent extrinsic properties from the target image. We further improve lighting transfer through a learned adaptor that injects the target's latent extrinsic properties via a cross-attention and light-weight fine-tuning. Unlike traditional ControlNet, which generates images with conditional maps from a single scene, LumiNet processes latent representations from two different images - preserving geometry and albedo from the source while transferring lighting characteristics from the target. Experiments demonstrate that our method successfully transfers complex lighting phenomena including specular highlights and indirect illumination across scenes with varying spatial layouts and materials, outperforming existing approaches on challenging indoor scenes using only images as input.

****

FedBiP: Heterogeneous One-Shot Federated Learning with Personalized Latent Diffusion Models

Haokun Chen, Hang Li, Yao Zhang, Jinhe Bi, Gengyuan Zhang, Yueqi Zhang, Philip Torr, Jindong Gu, Denis Krompass, Volker Tresp; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30440-30450

One-Shot Federated Learning (OSFL), a special decentralized machine learning paradigm, has recently gained significant attention. OSFL requires only a single round of client data or model upload, which reduces communication costs and mitiga

tes privacy threats compared to traditional FL. Despite these promising prospects, existing methods face challenges due to client data heterogeneity and limited data quantity when applied to real-world OSFL systems. Recently, Latent Diffusion Models (LDM) have shown remarkable advancements in synthesizing high-quality images through pretraining on large-scale datasets, thereby presenting a potential solution to overcome these issues. However, directly applying pretrained LDM to heterogeneous OSFL results in significant distribution shifts in synthetic data, leading to performance degradation in classification models trained on such data. This issue is particularly pronounced in rare domains, such as medical imaging, which are underrepresented in LDM's pretraining data. To address this challenge, we propose Federated Bi-Level Personalization (FedBiP), which personalizes the pretrained LDM at both instance-level and concept-level. Hereby, FedBiP synthesizes images following the client's local data distribution without compromising the privacy regulations. FedBiP is also the first approach to simultaneously address feature space heterogeneity and client data scarcity in OSFL. Our method is validated through extensive experiments on three OSFL benchmarks with feature space heterogeneity, as well as on challenging medical and satellite image datasets with label heterogeneity. The results demonstrate the effectiveness of FedBiP, which substantially outperforms other OSFL methods. Our code is available at \href https://github.com/HaokunChen245/FedBiP https://github.com/HaokunChen245/FedBiP .

****

DiC: Rethinking Conv3x3 Designs in Diffusion Models

Yuchuan Tian, Jing Han, Chengcheng Wang, Yuchen Liang, Chao Xu, Hanting Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2469-2478

Diffusion models have shown exceptional performance in visual generation tasks. Recently, these models have shifted from traditional U-Shaped CNN-Attention hybrid structures to fully transformer-based isotropic architectures. While these transformers exhibit strong scalability and performance, their reliance on complicated self-attention operation results in slow inference speeds. Contrary to these works, we rethink one of the simplest yet fastest module in deep learning, 3x3 Convolution, to construct a scaled-up purely convolutional diffusion model. We first discover that an Encoder-Decoder Hourglass design outperforms scalable isotropic architectures for Conv3x3, but still under-performing our expectation. Further improving the architecture, we introduce sparse skip connections to reduce redundancy and improve scalability. Based on the architecture, we introduce conditioning improvements including stage-specific embeddings, mid-block condition injection, and conditional gating. These improvements lead to our proposed Diffusion CNN (DiC), which serves as a swift yet competitive diffusion architecture baseline. Experiments on various scales and settings show that DiC surpasses existing diffusion transformers by considerable margins in terms of performance while keeping a good speed advantage. Project page: https://github.com/YuchuanTian/DiC

****

Dynamic Camera Poses and Where to Find Them

Chris Rockwell, Joseph Tung, Tsung-Yi Lin, Ming-Yu Liu, David F. Fouhey, Chen-Hsuan Lin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12444-12455

Annotating camera poses on dynamic Internet videos at scale is critical for advancing fields like realistic video generation and simulation. However, collecting such a dataset is difficult, as most Internet videos are unsuitable for pose estimation. Furthermore, annotating dynamic Internet videos present significant challenges even for state-of-the-art methods. In this paper, we introduce DynPose-100K, a large-scale dataset of dynamic Internet videos annotated with camera poses. Our collection pipeline addresses filtering using a carefully combined set of task-specific and generalist models. For pose estimation, we combine the latest techniques of point tracking, dynamic masking, and structure-from-motion to achieve improvements over the state-of-the-art approaches. Our analysis and experiments demonstrate that DynPose-100K is both large-scale and diverse across sever

al key attributes, opening up avenues for advancements in various downstream app
lications.
****

MoSca: Dynamic Gaussian Fusion from Casual Videos via 4D Motion Scaffolds
Jiahui Lei, Yijia Weng, Adam W. Harley, Leonidas Guibas, Kostas Daniilidis; Proc
eedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025,
pp. 6165-6177
We introduce 4D Motion Scaffolds (MoSca), a modern 4D reconstruction system desi
gned to reconstruct and synthesize novel views of dynamic scenes from monocular
videos captured casually in the wild. To address such a challenging and ill-pose
d inverse problem, we leverage prior knowledge from foundational vision models a
nd lift the video data to a novel Motion Scaffold (MoSca) representation, which
compactly and smoothly encodes the underlying motions/deformations. The scene ge
ometry and appearance are then disentangled from the deformation field and are e
ncoded by globally fusing the Gaussians anchored onto the MoSca and optimized vi
a Gaussian Splatting. Additionally, camera focal length and poses can be solved
using bundle adjustment without the need of any other pose estimation tools. Exp
eriments demonstrate state-of-the-art performance on dynamic rendering benchmark
s and its effectiveness on real videos.
****

GCE-Pose: Global Context Enhancement for Category-level Object Pose Estimation
Weihang Li, Hongli XU, Junwen Huang, Hyunjun Jung, Peter KT Yu, Nassir Navab, Be
njamin Busam; Proceedings of the Computer Vision and Pattern Recognition Confere
nce (CVPR), 2025, pp. 27154-27165
A key challenge in model-free category-level pose estimation is the extraction o
f contextual object features that generalize across varying instances within a s
pecific category. Recent approaches leverage foundational features to capture se
mantic and geometry cues from data. However, these approaches fail under partial
 visibility. We overcome this with a first-complete-then-aggregate strategy for
feature extraction utilizing class priors. In this paper, we present GCE-Pose, a
 method that enhances pose estimation for novel instances by integrating categor
y-level global context prior. GCE-Pose first performs semantic shape reconstruct
ion with a proposed Semantic Shape Reconstruction (SSR) module. Given an unseen
partial RGB-D object instance, our SSR module reconstructs the instance's global
 geometry and semantics by deforming category-specific 3D semantic prototypes th
rough a learned deep Linear Shape Model. We then introduce a Global Context Enha
nced (GCE) feature fusion module that effectively fuses features from partial RG
B-D observations and the reconstructed global context. Extensive experiments val
idate the impact of our global context prior and the effectiveness of the GCE fu
sion module, demonstrating that GCE-Pose significantly outperforms existing meth
ods on challenging real-world datasets HouseCat6D and NOCS-REAL275.
****

OmniGen: Unified Image Generation
Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Ch
aofan Li, Shuting Wang, Tiejun Huang, Zheng Liu; Proceedings of the Computer Vis
ion and Pattern Recognition Conference (CVPR), 2025, pp. 13294-13304
The emergence of Large Language Models (LLMs) has unified language generation ta
sks and revolutionized human-machine interaction. However, in the realm of image
 generation, a unified model capable of handling various tasks within a single f
ramework remains largely unexplored. In this work, we introduce OmniGen, a new d
iffusion model for unified image generation. OmniGen is characterized by the fol
lowing features: 1) Unification: OmniGen not only demonstrates text-to-image gen
eration capabilities but also inherently supports various downstream tasks, such
 as image editing, subject-driven generation, and visual-conditional generation.
 2) Simplicity: The architecture of OmniGen is highly simplified, eliminating th
e need for additional plugins. Moreover, compared to existing diffusion models,
it is more user-friendly and can complete complex tasks end-to-end through instr
uctions without the need for extra intermediate steps, greatly simplifying the i
mage generation workflow. 3) Knowledge Transfer: Benefit from learning in a unif
ied format, OmniGen effectively transfers knowledge across different tasks, mana

ges unseen tasks and domains, and exhibits novel capabilities. We also explore t
he model's reasoning capabilities and potential applications of the chain-of-tho
ught mechanism. This work represents the first attempt at a general-purpose imag
e generation model, and we will release our resources at https://github.com/Vect
orSpaceLab/OmniGen to foster future advancements.
****

QuCOOP: A Versatile Framework for Solving Composite and Binary-Parametrised Prob
lems on Quantum Annealers
Natacha Kuete Meli, Vladislav Golyanik, Marcel Seelbach Benkner, Michael Moeller
; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR),
2025, pp. 11395-11405
There is growing interest in solving computer vision problems such as mesh or po
int set alignment using Adiabatic Quantum Computing (AQC). Unfortunately, modern
 experimental AQC devices such as D-Wave only support Quadratic Unconstrained Bi
nary Optimisation (QUBO) problems, which severely limits their applicability. Th
is paper proposes a new way to overcome this limitation and introduces QuCOOP, a
n optimisation framework extending the scope of AQC to composite and binary-para
metrised, possibly non-quadratic problems. The key idea of QuCOOP is to iterativ
ely approximate the original objective function by a sequel of local (intermedia
te) QUBO forms, whose binary parameters can be sampled on AQC devices. We experi
ment with quadratic assignment problems, shape matching and point set registrati
on without knowing the correspondences in advance. Our approach achieves state-o
f-the-art results across multiple instances of tested problems.
****

Mesh Mamba: A Unified State Space Model for Saliency Prediction in Non-Textured
and Textured Meshes
Kaiwei Zhang, Dandan Zhu, Xiongkuo Min, Guangtao Zhai; Proceedings of the Comput
er Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16219-16228
Mesh saliency enhances the adaptability of 3D vision by identifying and emphasiz
ing regions that naturally attract visual attention. To investigate the interact
ion between geometric structure and texture in shaping visual attention, we esta
blish a comprehensive mesh saliency dataset, which is the first to systematicall
y capture the differences in saliency distribution under both textured and non-t
extured visual conditions. Furthermore, we introduce mesh Mamba, a unified salie
ncy prediction model based on a state space model (SSM), designed to adapt acros
s various mesh types. Mesh Mamba effectively analyzes the geometric structure of
 the mesh while seamlessly incorporating texture features into the topological f
ramework, ensuring coherence throughout appearance-enhanced modeling. More impor
tantly, by subgraph embedding and a bidirectional SSM, the model enables global
context modeling for both local geometry and texture, preserving the topological
 structure and improving the understanding of visual details and structural comp
lexity. Through extensive theoretical and empirical validation, our model not on
ly improves performance across various mesh types but also demonstrates high sca
lability and versatility, particularly through cross validations of various visu
al features.
****

Not Just Text: Uncovering Vision Modality Typographic Threats in Image Generatio
n Models
Hao Cheng, Erjia Xiao, Jiayan Yang, Jiahang Cao, Qiang Zhang, Jize Zhang, Kaidi
Xu, Jindong Gu, Renjing Xu; Proceedings of the Computer Vision and Pattern Recog
nition Conference (CVPR), 2025, pp. 2997-3007
Current image generation models can effortlessly produce high-quality, highly re
alistic images, but this also increases the risk of misuse. In various Text-to-I
mage or Image-to-Image tasks, attackers can generate a series of images containi
ng inappropriate content by simply editing the language modality input. To mitig
ate this security concern, numerous guarding or defensive strategies have been p
roposed, with a particular emphasis on safeguarding language modality. However,
in practical applications, threats in the vision modality, particularly in tasks
 involving the editing of real-world images, present heightened security risks a
s they can easily infringe upon the rights of the image owner. Therefore, this

paper employs a method named typographic attack to reveal that various image gen
eration models are also susceptible to threats within the vision modality. Furth
ermore, we also evaluate the defense performance of various existing methods whe
n facing threats in the vision modality and uncover their ineffectiveness. Final
ly, we propose the Vision Modal Threats in Image Generation Models (VMT-IGMs) da
taset, which would serve as a baseline for evaluating the vision modality vulner
ability of various image generation models.
****

SILMM: Self-Improving Large Multimodal Models for Compositional Text-to-Image Ge
neration
Leigang Qu, Haochuan Li, Wenjie Wang, Xiang Liu, Juncheng Li, Liqiang Nie, Tat-S
eng Chua; Proceedings of the Computer Vision and Pattern Recognition Conference
(CVPR), 2025, pp. 18497-18508
Large Multimodal Models (LMMs) have demonstrated impressive capabilities in mult
imodal understanding and generation, pushing forward advancements in text-to-ima
ge generation.However, achieving accurate text-image alignment for LMMs, particu
larly in compositional scenarios, remains challenging. Existing approaches, such
 as layout planning for multi-step generation and learning from human feedback o
r AI feedback, depend heavily on prompt engineering, costly human annotations, a
nd continual upgrading, limiting flexibility and scalability. In this work, we i
ntroduce a model-agnostic iterative self-improvement framework (**SILMM**) that
can enable LMMs to provide helpful and scalable self-feedback and optimize text-
image alignment via Direct Preference Optimization (DPO). DPO can readily applie
d to LMMs that use discrete visual tokens as intermediate image representations;
 while it is less suitable for LMMs with continuous visual features, as obtainin
g generation probabilities is challenging.To adapt SILMM to LMMs with continuous
 features, we propose a diversity mechanism to obtain diverse representations an
d a kernel-based continuous DPO for alignment. Extensive experiments on three co
mpositional text-to-image generation benchmarks validate the effectiveness and s
uperiority of SILMM, showing improvements exceeding 30% on T2I-CompBench++ and a
round 20% on DPG-Bench.
****

Calibrated Multi-Preference Optimization for Aligning Diffusion Models
Kyungmin Lee, Xiahong Li, Qifei Wang, Junfeng He, Junjie Ke, Ming-Hsuan Yang, Ir
fan Essa, Jinwoo Shin, Feng Yang, Yinxiao Li; Proceedings of the Computer Vision
 and Pattern Recognition Conference (CVPR), 2025, pp. 18465-18475
Aligning text-to-image (T2I) diffusion models with prefer-ence optimization is v
aluable for human-annotated datasets, but the heavy cost of manual data collecti
on limits scalability. Using reward models offers an alternative, however, curre
nt preference optimization methods fall short in exploiting the rich information
, as they only consider pairwise preference distribution. Furthermore, they lack
 generalization to multi-preference scenarios and struggle to handle inconsisten
cies between rewards. To address this, we present Calibrated Preference Optimiza
tion (CaPO), a novel method to align T2I diffusion models by incorporating the g
eneral preference from multiple reward models without human annotated data. The
core of our approach involves a reward calibration method to approximate the gen
eral preference by computing the expected win-rate against the samples generated
 by the pretrained models. Additionally, we propose a frontier-based pair select
ion method that effectively manages the multi-preference distribution by selecti
ng pairs from Pareto frontiers. Finally, we use regression loss to fine-tune dif
fusion models to match the difference between calibrated rewards of a selected p
air. Experimental results show that CaPO consistently outperforms prior methods,
 such as Direct Preference Optimization (DPO), in both single and multi-reward s
ettings validated by evaluation on T2I benchmarks, including GenEval and T2I-Com
pbench.
****

Learning from Neighbors: Category Extrapolation for Long-Tail Learning
Shizhen Zhao, Xin Wen, Jiahui Liu, Chuofan Ma, Chunfeng Yuan, Xiaojuan Qi; Proce
edings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, p
p. 30483-30492

Balancing training on long-tail data distributions remains a long-standing challenge in deep learning. While methods such as re-weighting and re-sampling help alleviate the imbalance issue, limited sample diversity continues to hinder models from learning robust and generalizable feature representations, particularly for tail classes. In contrast to existing methods, we offer a novel perspective on long-tail learning, inspired by an observation: datasets with finer granularity tend to be less affected by data imbalance. In this paper, we investigate this phenomenon through both quantitative and qualitative studies, showing that increased granularity enhances the generalization of learned features in tail categories. Motivated by these findings, we propose a method to increase dataset granularity through category extrapolation. Specifically, we introduce open-set fine-grained classes that are related to existing ones, aiming to enhance representation learning for both head and tail classes. To automate the curation of auxiliary data, we leverage large language models (LLMs) as knowledge bases to search for auxiliary categories and retrieve relevant images through web crawling. To prevent the overwhelming presence of auxiliary classes from disrupting training, we introduce a neighbor-silencing loss that encourages the model to focus on class discrimination within the target dataset. During inference, the classifier weights for auxiliary categories are masked out, leaving only the target class weights for use. Extensive experiments on three standard long-tail benchmarks demonstrate the effectiveness of our approach, notably outperforming strong baseline methods that use the same amount of data.
****

Material Anything: Generating Materials for Any 3D Object via Diffusion
Xin Huang, Tengfei Wang, Ziwei Liu, Qing Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26556-26565
We present **Material Anything**, a fully-automated, unified diffusion framework designed to generate physically-based materials for 3D objects. Unlike existing methods that rely on complex pipelines or case-specific optimizations, Material Anything offers a robust, end-to-end solution adaptable to objects under diverse lighting conditions. Our approach leverages a pre-trained image diffusion model, enhanced with a triple-head architecture and rendering loss to improve stability and material quality. Additionally, we introduce confidence masks as a dynamic switcher within the diffusion model, enabling it to effectively handle both textured and texture-less objects across varying lighting conditions. By employing a progressive material generation strategy guided by these confidence masks, along with a UV-space material refiner, our method ensures consistent, UV-ready material outputs. Extensive experiments demonstrate our approach outperforms existing methods across a wide range of object categories and lighting conditions.
****

TokenHSI: Unified Synthesis of Physical Human-Scene Interactions through Task Tokenization
Liang Pan, Zeshi Yang, Zhiyang Dou, Wenjia Wang, Buzhen Huang, Bo Dai, Taku Komura, Jingbo Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5379-5391
Synthesizing diverse and physically plausible Human-Scene Interactions (HSI) is pivotal for both computer animation and embodied AI. Despite encouraging progress, current methods mainly focus on developing separate controllers, each specialized for a specific interaction task. This significantly hinders the ability to tackle a wide variety of challenging HSI tasks that require the integration of multiple skills, e.g., sitting down while carrying an object (see Fig. 1). To address this issue, we present TokenHSI, a single, unified transformer-based policy capable of multi-skill unification and flexible adaptation. The key insight is to model the humanoid proprioception as a separate shared token and combine it with distinct task tokens via a masking mechanism. Such a unified policy enables effective knowledge sharing across skills, thereby facilitating the multi-task training. Moreover, our policy architecture supports variable length inputs, enabling flexible adaptation of learned skills to new scenarios. By training additional task tokenizers, we can not only modify the geometries of interaction targets but also coordinate multiple skills to address complex tasks. The experiments

demonstrate that our approach can significantly improve versatility, adaptabilit
y, and extensibility in various HSI tasks.
****

ShapeShifter: 3D Variations Using Multiscale and Sparse Point-Voxel Diffusion
Nissim Maruani, Wang Yifan, Matthew Fisher, Pierre Alliez, Mathieu Desbrun; Proc
eedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025,
pp. 605-617
This paper proposes ShapeShifter, a new 3D generative model that learns to synth
esize shape variations based on a single reference model. While generative metho
ds for 3D objects have recently attracted much attention, current techniques oft
en lack geometric details and/or require long training times and large resources
. Our approach remedies these issues by combining sparse voxel grids and multisc
ale point, normal, and color sampling within an encoder-free neural architecture
 that can be trained efficiently and in parallel. We show that our resulting var
iations better capture the fine details of their original input and can capture
more general types of surfaces than previous SDF-based methods. Moreover, we off
er interactive generation of 3D shape variants, allowing more human control in t
he design loop if needed.
****

ImagineFSL: Self-Supervised Pretraining Matters on Imagined Base Set for VLM-bas
ed Few-shot Learning
Haoyuan Yang, Xiaoou Li, Jiaming Lv, Xianjun Cheng, Qilong Wang, Peihua Li; Proc
eedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025,
pp. 30020-30031
Adapting CLIP models for few-shot recognition has recently attracted significant
 attention. Despite considerable progress, these adaptations remain hindered by
the pervasive challenge of data scarcity. Text-to-image models, capable of gener
ating abundant photorealistic labeled images, offer a promising solution. Howeve
r, existing approaches treat synthetic images merely as complements to real imag
es, rather than as standalone knowledge repositories stemming from distinct foun
dation models. To overcome this limitation, we reconceptualize synthetic images
as an *imagined base set*, i.e., a unique, large-scale synthetic dataset encompa
ssing diverse concepts. We introduce a novel CLIP adaptation methodology called
ImagineFSL, involving pretraining on the imagined base set followed by fine-tu
ning on downstream few-shot tasks. We find that, compared to no pretraining, bot
h supervised and self-supervised pretraining are beneficial, with the latter p
roviding better performance. Building on this finding, we propose an improved se
lf-supervised method tailored for few-shot scenarios, enhancing the transferabil
ity of representations from synthetic to real image domains. Additionally, we pr
esent an image generation pipeline that employs chain-of-thought and in-context
learning techniques, harnessing foundation models to automatically generate dive
rse, realistic images. Our methods are validated across eleven datasets, consist
ently outperforming state-of-the-art methods by substantial margins.
****

Continuous Locomotive Crowd Behavior Generation
Inhwan Bae, Junoh Lee, Hae-Gon Jeon; Proceedings of the Computer Vision and Patt
ern Recognition Conference (CVPR), 2025, pp. 22416-22431
Modeling and reproducing crowd behaviors are important in various domains includ
ing psychology, robotics, transport engineering and virtual environments. Conven
tional methods have focused on synthesizing momentary scenes, which have difficu
lty in replicating the continuous nature of real-world crowds. In this paper, we
 introduce a novel method for automatically generating continuous, realistic cro
wd trajectories with heterogeneous behaviors and interactions among individuals.
 We first design a crowd emitter model. To do this, we obtain spatial layouts fr
om single input images, including a segmentation map, appearance map, population
 density map and population probability, prior to crowd generation. The emitter
then continually places individuals on the timeline by assigning independent beh
avior characteristics such as agents' type, pace, and start/end positions using
diffusion models. Next, our crowd simulator produces their long-term locomotions
. To simulate diverse actions, it can augment their behaviors based on a Markov

chain. As a result, our overall framework populates the scenes with heterogeneous crowd behaviors by alternating between the proposed emitter and simulator. Note that all the components in the proposed framework are user-controllable. Lastly, we propose a benchmark protocol to evaluate the realism and quality of the generated crowds in terms of the scene-level population dynamics and the individual-level trajectory accuracy. We demonstrate that our approach effectively models diverse crowd behavior patterns and generalizes well across different geographical environments. Code is publicly available at https://github.com/InhwanBae/CrowdES.
****

Project-Probe-Aggregate: Efficient Fine-Tuning for Group Robustness
Beier Zhu, Jiequan Cui, Hanwang Zhang, Chi Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25487-25496
While image-text foundation models have succeeded across diverse downstream tasks, they still face challenges in the presence of spurious correlations between the input and label. To address this issue, we propose a simple three-step approach-Project-Probe-Aggregate (PPA)-that enables parameter-efficient fine-tuning for foundation models without relying on group annotations. Building upon the failure-based debiasing scheme, our method, PPA, improves its two key components: minority samples identification and the robust training algorithm. Specifically, we first train biased classifiers by projecting image features onto the nullspace of class proxies from text encoders. Next, we infer group labels using the biased classifier and probe group targets with prior correction. Finally, we aggregate group weights of each class to produce the debiased classifier. Our theoretical analysis shows that our PPA enhances minority group identification and is Bayes optimal for minimizing the balanced group error, mitigating spurious correlations. Extensive experimental results confirm the effectiveness of our PPA: it outperforms the state-of-the-art by an average worst-group accuracy while requiring less than 0.01% tunable parameters without training group labels.
****

Implicit Bias Injection Attacks against Text-to-Image Diffusion Models
Huayang Huang, Xiangye Jin, Jiaxu Miao, Yu Wu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28779-28789
The proliferation of text-to-image diffusion models (T2I DMs) has led to an increased presence of AI-generated images in daily life. However, biased T2I models can generate content with specific tendencies, potentially influencing people's perceptions. Intentional exploitation of these biases risks conveying misleading information to the public. Current research on bias primarily addresses explicit biases with recognizable visual patterns, such as skin color and gender. This paper introduces a novel form of implicit bias that lacks explicit visual features but can manifest in diverse ways across various semantic contexts. This subtle and versatile nature makes this bias challenging to detect, easy to propagate, and adaptable to a wide range of scenarios. We further propose an implicit bias injection attack framework (IBI-Attacks) against T2I diffusion models by precomputing a general bias direction in the prompt embedding space and adaptively adjusting it based on different inputs. Our attack module can be seamlessly integrated into pre-trained diffusion models in a plug-and-play manner without direct manipulation of user input or model retraining. Extensive experiments validate the effectiveness of our scheme in introducing bias through subtle and diverse modifications while preserving the original semantics.The strong concealment and transferability of our attack across various scenarios further underscore the significance of our approach.
****

ROICtrl: Boosting Instance Control for Visual Generation
Yuchao Gu, Yipin Zhou, Yunfan Ye, Yixin Nie, Licheng Yu, Pingchuan Ma, Kevin Qinghong Lin, Mike Zheng Shou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23658-23667
Natural language often struggles to accurately associate positional and attribute information with multiple instances, which limits current text-based visual generation models to simpler compositions featuring only a few dominant instances.

To address this limitation, this work enhances diffusion models by introducing regional instance control, where each instance is governed by a bounding box paired with a free-form caption. Previous methods in this area typically rely on implicit position encoding or explicit attention masks to separate regions of interest (ROIs), resulting in either inaccurate coordinate injection or large computational overhead. Inspired by ROI-Align in object detection, we introduce a complementary operation called ROI-Unpool. Together, ROI-Align and ROI-Unpool enable explicit, efficient, and accurate ROI manipulation on high-resolution feature maps for visual generation. Building on ROI-Unpool, we propose ROICtrl, an adapter for pretrained diffusion models that enables precise regional instance control. ROICtrl is compatible with community-finetuned diffusion models, as well as with existing spatial-based add-ons (e.g., ControlNet, T2I-Adapter) and embedding-based add-ons (e.g., IP-Adapter, ED-LoRA), extending their applications to multi-instance generation. Experiments show that ROICtrl achieves superior performance in regional instance control while significantly reducing computational costs.

****

FRESA: Feedforward Reconstruction of Personalized Skinned Avatars from Few Images

Rong Wang, Fabian Prada, Ziyan Wang, Zhongshi Jiang, Chengxiang Yin, Junxuan Li, Shunsuke Saito, Igor Santesteban, Javier Romero, Rohan Joshi, Hongdong Li, Jason Saragih, Yaser Sheikh; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 281-291

We present a novel method for reconstructing personalized 3D human avatars with realistic animation from only a few images. Due to the large variations in body shapes, poses, and cloth types, existing methods mostly require hours of per-subject optimization during inference, which limits their practical applications. In contrast, we learn a universal prior from over a thousand clothed humans to achieve instant feedforward generation and zero-shot generalization. Specifically, instead of rigging the avatar with shared skinning weights, we jointly infer personalized avatar shape, skinning weights, and pose-dependent deformations, which effectively improves overall geometric fidelity and reduces deformation artifacts. Moreover, to normalize pose variations and resolve coupled ambiguity between canonical shapes and skinning weights, we design a 3D canonicalization process to produce pixel-aligned initial conditions, which helps to reconstruct fine-grained geometric details. We then propose a multi-frame feature aggregation to robustly reduce artifacts introduced in canonicalization and fuse a plausible avatar preserving person-specific identities. Finally, we train the model in an end-to-end framework on a large-scale capture dataset, which contains diverse human subjects paired with high-quality 3D scans. Extensive experiments show that our method generates more authentic reconstruction and animation than state-of-the-arts, and can be directly generalized to inputs from casually taken phone photos. Project page and code is available at https://github.com/rongakowang/FRESA.

****

ReasonGrounder: LVLM-Guided Hierarchical Feature Splatting for Open-Vocabulary 3D Visual Grounding and Reasoning

Zhenyang Liu, Yikai Wang, Sixiao Zheng, Tongying Pan, Longfei Liang, Yanwei Fu, Xiangyang Xue; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3718-3727

Open-vocabulary 3D visual grounding and reasoning aim to localize objects in a scene based on implicit language descriptions, even when they are occluded. This ability is crucial for tasks such as vision-language navigation and autonomous robotics. However, current methods struggle because they rely heavily on fine-tuning with 3D annotations and mask proposals, which limits their ability to handle diverse semantics and common knowledge required for effective reasoning. To address this, we propose ReasonGrounder, an LVLM-guided framework that uses hierarchical 3D feature Gaussian fields for adaptive grouping based on physical scale, enabling open-vocabulary 3D grounding and reasoning. ReasonGrounder interprets implicit instructions using large vision-language models (LVLM) and localizes occluded objects through 3D Gaussian splatting. By incorporating 2D segmentation masks from the Segment Anything Model (SAM) and multi-view CLIP embeddings, Reason

Grounder selects Gaussian groups based on object scale, enabling accurate localization through both explicit and implicit language understanding, even in novel, occluded views. We also contribute ReasoningGD, a new dataset containing over 10K scenes and 2 million annotations for evaluating open-vocabulary 3D grounding and amodal perception under occlusion. Experiments show that ReasonGrounder significantly improves 3D grounding accuracy in real-world scenarios.

****

Cropper: Vision-Language Model for Image Cropping through In-Context Learning
Seung Hyun Lee, Jijun Jiang, Yiran Xu, Zhuofang Li, Junjie Ke, Yinxiao Li, Junfeng He, Steven Hickson, Katie Datsenko, Sangpil Kim, Ming-Hsuan Yang, Irfan Essa, Feng Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30010-30019

The goal of image cropping is to identify visually appealing crops in an image. Conventional methods are trained on specific datasets and fail to adapt to new requirements. Recent breakthroughs in large vision-language models (VLMs) enable visual in-context learning without explicit training. However, downstream tasks with VLMs remain under explored. In this paper, we propose an effective approach to leverage VLMs for image cropping. First, we propose an efficient prompt retrieval mechanism for image cropping to automate the selection of in-context examples. Second, we introduce an iterative refinement strategy to iteratively enhance the predicted crops. The proposed framework, we refer to as Cropper, is applicable to a wide range of cropping tasks, including free-form cropping, subject-aware cropping, and aspect ratio-aware cropping. Extensive experiments demonstrate that Cropper significantly outperforms state-of-the-art methods across several benchmarks.

****

Advancing Adversarial Robustness in GNeRFs: The IL2-NeRF Attack
Nicole Meng, Caleb Manicke, Ronak Sahu, Caiwen Ding, Yingjie Lao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16388-16397

Generalizable Neural Radiance Fields (GNeRF) are recognized as one of the most promising techniques for novel view synthesis and 3D model generation in real-world applications. However, like other generative models in computer vision, ensuring their adversarial robustness against various threat models is essential for practical use. The pioneering work in this area, NeRFool, introduced a state-of-the-art attack that targets GNeRFs by manipulating source views before feature extraction, successfully disrupting the color and density results of the constructed views. Building on this foundation, we propose IL2-NeRF (Iterative L2 NeRF Attack), a novel adversarial attack method that explores a new threat model (in the L2 domain) for attacking GNeRFs. We evaluated IL2-NeRF against two standard GNeRF models across three benchmark datasets, demonstrating similar performance compared to NeRFool, based on the same evaluation metrics proposed by NeRFool. Our results establish IL2-NeRF as the first adversarial method for GNeRFs under the L2 norm. We establish a foundational L2 threat model for future research, enabling direct performance comparisons while introducing a smoother, image-wide perturbation approach in Adversarial 3D Reconstruction. Our code is available at: https://github.com/The-NRC-SCAR-Group/IL2-NeRF

****

WonderWorld: Interactive 3D Scene Generation from a Single Image
Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T. Freeman, Jiajun Wu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5916-5926

We present WonderWorld, a novel framework for interactive 3D scene generation that enables users to interactively specify scene contents and layout and see the created scenes in low latency. The major challenge lies in achieving fast generation of 3D scenes. Existing scene generation approaches fall short of speed as they often require (1) progressively generating many views and depth maps, and (2) time-consuming optimization of the scene representations. Our approach does not need multiple views, and it leverages a geometry-based initialization that significantly reduces optimization time. Another challenge is generating coherent g

eometry that allows all scenes to be connected. We introduce the guided depth di
ffusion that allows partial conditioning of depth estimation. WonderWorld genera
tes connected and diverse 3D scenes in less than 10 seconds on a single A6000 GP
U, enabling real-time user interaction and exploration. Our interactive demo, fu
ll code, data, and software can be found at https://kovenyu.com/WonderWorld/
****

A Lightweight UDF Learning Framework for 3D Reconstruction Based on Local Shape Functions

Jiangbei Hu, Yanggeng Li, Fei Hou, Junhui Hou, Zhebin Zhang, Shengfa Wang, Na Le
i, Ying He; Proceedings of the Computer Vision and Pattern Recognition Conferenc
e (CVPR), 2025, pp. 1297-1307

Unsigned distance fields (UDFs) provide a versatile framework for representing a
 diverse array of 3D shapes, encompassing both watertight and non-watertight geo
metries. Traditional UDF learning methods typically require extensive training o
n large 3D shape datasets, which is costly and necessitates re-training for new
datasets. This paper presents a novel neural framework, *LoSF-UDF*, for reconstr
ucting surfaces from 3D point clouds by leveraging local shape functions to lear
n UDFs. We observe that 3D shapes manifest simple patterns in localized regions,
 prompting us to develop a training dataset of point cloud patches characterized
 by mathematical functions that represent a continuum from smooth surfaces to sh
arp edges and corners. Our approach learns features within a specific radius aro
und each query point and utilizes an attention mechanism to focus on the crucial
 features for UDF estimation. Despite being highly lightweight, with only 653 KB
 of trainable parameters and a modest-sized training dataset with 0.5 GB storage
, our method enables efficient and robust surface reconstruction from point clou
ds without requiring for shape-specific training. Furthermore, our method exhibi
ts enhanced resilience to noise and outliers in point clouds compared to existin
g methods. We conduct comprehensive experiments and comparisons across various d
atasets, including synthetic and real-scanned point clouds, to validate our meth
od's efficacy. Notably, our lightweight framework offers rapid and reliable init
ialization for other unsupervised iterative approaches, improving both the effic
iency and accuracy of their reconstructions.
****

DiffCAM: Data-Driven Saliency Maps by Capturing Feature Differences

Xingjian Li, Qiming Zhao, Neelesh Bisht, Mostofa Rafid Uddin, Jin Yu Kim, Bryan
Zhang, Min Xu; Proceedings of the Computer Vision and Pattern Recognition Confer
ence (CVPR), 2025, pp. 10327-10337

In recent years, the interpretability of Deep Neural Networks (DNNs) has garnere
d significant attention, particularly due to their widespread deployment in crit
ical domains like healthcare, finance, and autonomous systems. To address the ch
allenge of understanding how DNNs make decisions, Explainable AI (XAI) methods,
such as saliency maps, have been developed to provide insights into the inner wo
rkings of these models. This paper introduces DiffCAM, a novel XAI method design
ed to overcome limitations in existing Class Activation Map (CAM)-based techniqu
es, which often rely on decision boundary gradients to estimate feature importan
ce. DiffCAM differentiates itself by considering the actual data distribution of
 the reference class, identifying feature importance based on how a target examp
le differs from reference examples. This approach captures the most discriminati
ve features without relying on decision boundaries or prediction results, making
 DiffCAM applicable to a broader range of models, including foundation models. T
hrough extensive experiments, we demonstrate the superior performance and flexib
ility of DiffCAM in providing meaningful explanations across diverse datasets an
d scenarios.
****

PolarNeXt: Rethink Instance Segmentation with Polar Representation

Jiacheng Sun, Xinghong Zhou, Yiqiang Wu, Bin Zhu, Jiaxuan Lu, Yu Qin, Xiaomao Li
; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR),
2025, pp. 19315-19324

One of the roadblocks for instance segmentation today is heavy computational ove
rhead and model parameters. Previous methods based on Polar Representation made

the initial mark to address this challenge by formulating instance segmentation as polygon detection, but failed to align with mainstream methods in performance. In this paper, we highlight that Representation Errors, arising from the limited capacity of polygons to capture boundary details, have long been overlooked, which results in severe performance degradation. Observing that optimal starting point selection effectively alleviates this issue, we propose an Adaptive Polygonal Sample Decision strategy to dynamically capture the positional variation of representation errors across samples. Additionally, we design a Union-aligned Rasterization Module to incorporate these errors into polygonal assessment, further advancing the proposed strategy. With these components, our framework called PolarNeXt achieves a remarkable performance boost of over 4.8% AP compared to other polar-based methods. PolarNeXt is markedly more lightweight and efficient than state-of-the-art instance segmentation methods, while achieving comparable segmentation accuracy. We expect this work will open up a new direction for instance segmentation in high-resolution images and resource-limited scenarios.

****

ScaMo: Exploring the Scaling Law in Autoregressive Motion Generation Model

Shunlin Lu, Jingbo Wang, Zeyu Lu, Ling-Hao Chen, Wenxun Dai, Junting Dong, Zhiyang Dou, Bo Dai, Ruimao Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27872-27882

The scaling law has been validated in various domains, such as natural language processing (NLP) and massive computer vision tasks; however, its application to motion generation remains largely unexplored. In this paper, we introduce a scalable motion generation framework that includes the motion tokenizer Motion FSQ-VAE and a text-prefix autoregressive transformer. Through comprehensive experiments, we observe the scaling behavior of this system. For the first time, we confirm the existence of scaling laws within the context of motion generation. Specifically, our results demonstrate that the normalized test loss of our prefix autoregressive models adheres to a logarithmic law in relation to compute budgets. Furthermore, we also confirm the power law between Non-Vocabulary Parameters, Vocabulary Parameters, and Data Tokens with respect to compute budgets respectively. Leveraging the scaling law, we predict the optimal transformer size, vocabulary size, and data requirements for a compute budget of 1e18. The test loss of the system, when trained with the optimal model size, vocabulary size, and required data, aligns precisely with the predicted test loss.

****

From Sparse Signal to Smooth Motion: Real-Time Motion Generation with Rolling Prediction Models

German Barquero, Nadine Bertsch, Manojkumar Marramreddy, Carlos Chacón, Filippo Arcadu, Ferran Rigual, Nicky Sijia He, Cristina Palmero, Sergio Escalera, Yuting Ye, Robin Kips; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1850-1860

In extended reality (XR), generating full-body motion of the users is important to understand their actions, drive their virtual avatars for social interaction, and convey a realistic sense of presence. While prior works focused on spatially sparse and always-on input signals from motion controllers, many XR applications opt for vision-based hand tracking for reduced user friction and better immersion. Compared to controllers, hand tracking signals are less accurate and can even be missing for an extended period of time. To handle such unreliable inputs, we present Rolling Prediction Model (RPM), an online and real-time approach that generates smooth full-body motion from temporally and spatially sparse input signals. Our model generates 1) accurate motion that matches the inputs (i.e., tracking mode) and 2) plausible motion when inputs are missing (i.e., synthesis mode). More importantly, RPM generates seamless transitions from tracking to synthesis, and vice versa. To demonstrate the practical importance of handling noisy and missing inputs, we present GORP, the first dataset of realistic sparse inputs from a commercial virtual reality (VR) headset with paired high quality body motion ground truth. GORP provides >14 hours of VR gameplay data from 28 people using motion controllers (spatially sparse) and hand tracking (spatially and temporally sparse). We benchmark RPM against the state of the art on both synthetic

data and GORP to highlight how we can bridge the gap for real-world applications with a realistic dataset and by handling unreliable input signals. Our code, pretrained models, and GORP dataset are available in the project webpage.
****

Imagine and Seek: Improving Composed Image Retrieval with an Imagined Proxy
You Li, Fan Ma, Yi Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3984-3993
The Zero-shot Composed Image Retrieval (ZSCIR) requires retrieving images that match the query image and the relative captions.Current methods focus on projecting the query image into the text feature space, subsequently combining them with features of query texts for retrieval. However, retrieving images only with the text features cannot guarantee detailed alignment due to the natural gap between images and text.In this paper, we introduce Imagined Proxy for CIR, a training-free method that creates a proxy image aligned with the query image and text description, enhancing query representation in the retrieval process.We first leverage the large language model's generalization capability to generate an image layout, and then apply both the query text and image for conditional generation.The robust query features are enhanced by merging the proxy image, query image, and text semantic perturbation. Our newly proposed balancing metric integrates text-based and proxy retrieval similarities, allowing for more accurate retrieval of the target image while incorporating image-side information into the process. Experiments on three public datasets demonstrate that our method significantly improves retrieval performances. We achieve state-of-the-art (SOTA) results on the CIRR dataset with a Recall@K of 70.07 at K=10. Additionally, we achieved an improvement in Recall@10 on the FashionIQ dataset, rising from 45.11 to 45.74, and improved the baseline performance in CIRCO with a mAPK@10 score, increasing from 32.24 to 34.26.
****

EMOVA: Empowering Language Models to See, Hear and Speak with Vivid Emotions
Kai Chen, Yunhao Gou, Runhui Huang, Zhili Liu, Daxin Tan, Jing Xu, Chunwei Wang, Yi Zhu, Yihan Zeng, Kuo Yang, Dingdong Wang, Kun Xiang, Haoyuan Li, Haoli Bai, Jianhua Han, Xiaohui Li, Weike Jin, Nian Xie, Yu Zhang, James T. Kwok, Hengshuang Zhao, Xiaodan Liang, Dit-Yan Yeung, Xiao Chen, Zhenguo Li, Wei Zhang, Qun Liu, Lanqing Hong, Lu Hou, Hang Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5455-5466
GPT-4o, an omni-modal model that enables vocal conversations with diverse emotions and tones, marks a milestone for omni-modal foundation models. However, empowering Large Language Models to perceive and generate images, texts, and speeches end-to-end with publicly available data remains challenging for the open-source community. Existing vision-language models rely on external tools for speech processing, while speech-language models still suffer from limited or totally without vision-understanding capabilities. To address this gap, we propose the EMOVA (EMotionally Omni-present Voice Assistant), to enable Large Language Models with end-to-end speech abilities while maintaining the leading vision-language performance. With a semantic-acoustic disentangled speech tokenizer, we surprisingly notice that omni-modal alignment can further enhance vision-language and speech abilities compared with the bi-modal aligned counterparts. Moreover, a lightweight style module is introduced for the flexible speech style controls including emotions and pitches. For the first time, EMOVA achieves state-of-the-art performance on both the vision-language and speech benchmarks, and meanwhile, supporting omni-modal spoken dialogue with vivid emotions.
****

SAM-REF: Introducing Image-Prompt Synergy during Interaction for Detail Enhancement in the Segment Anything Model
Chongkai Yu, Ting Liu, Anqi Li, Xiaochao Qu, Chengjing Wu, Luoqi Liu, Xiaolin Hu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19356-19365
Interactive segmentation is to segment the mask of the target object according to the user's interactive prompts. There are two mainstream strategies: early fusion and late fusion. Current specialist models utilize the early fusion strategy

that encodes the combination of images and prompts to target the prompted objects, yet repetitive complex computations on the images result in high latency. Late fusion models extract image embeddings once and merge them with the prompts in later interactions. This strategy avoids redundant image feature extraction and improves efficiency significantly. A recent milestone is the Segment Anything Model (SAM). However, this strategy limits the models ability to extract detailed information from the prompted target zone. To address this issue, we propose SAM-REF, a two-stage refinement framework that fully integrates images and prompts by using a lightweight refiner into the interaction of late fusion, which combines the accuracy of early fusion and maintains the efficiency of late fusion. Through extensive experiments, we show that our SAM-REF model outperforms the current state-of-the-art method in most metrics on segmentation quality without compromising efficiency.

****

DarkIR: Robust Low-Light Image Restoration

Daniel Feijoo, Juan C. Benito, Alvaro Garcia, Marcos V. Conde; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10879-10889

Photography during night or in dark conditions typically suffers from noise, low light and blurring issues due to the dim environment and the common use of long exposure. Although Deblurring and Low-light Image Enhancement (LLIE) are related under these conditions, most approaches in image restoration solve these tasks separately. In this paper, we present an efficient and robust neural network for multi-task low-light image restoration. Instead of following the current tendency of Transformer-based models, we propose new attention mechanisms to enhance the receptive field of efficient CNNs. Our method reduces the computational costs in terms of parameters and MAC operations compared to previous methods. Our model, DarkIR, achieves new state-of-the-art results on the popular LOLBlur, LOLv2 and Real-LOLBlur datasets, being able to generalize on real-world night images.

****

R2C: Mapping Room to Chessboard to Unlock LLM As Low-Level Action Planner

Ziyi Bai, Hanxuan Li, Bin Fu, Chuyan Xiong, Ruiping Wang, Xilin Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19456-19466

This paper explores using large language models (LLMs) as low-level action planners for embodied tasks. While LLMs excel as the robot's "brain" for high-level planning, they face challenges in directly controlling the "body" by generating precise low-level actions. This limitation arises from LLMs' strength in high-level conceptual understanding but their inability to handle spatial perception effectively, restricting their potential in embodied tasks. To address this, we bridge the gap by enabling LLMs to not only comprehend complex instructions but also produce actionable, low-level plans. We introduce Room to Chessboard (R2C), a novel semantic representation that maps environmental states onto a grid-based chessboard, empowering LLMs to generate specific low-level coordinates and guide the robot in a manner akin to playing a game of chess. To further enhance decision-making, we propose the Chain-of-Thought Decision (CoT-D) paradigm, which improves LLMs' interpretability and context-awareness in spatial reasoning. By jointly training LLMs for high-level task decomposition and low-level action generation, we create a unified "brain-body" system capable of handling complex, free-form instructions while producing precise low-level actions, allowing the robot to flexibly control its movements and adapt to varying tasks. We validate R2C using both fine-tuned open-source LLMs and GPT-4, demonstrating effectiveness on the challenging ALFRED benchmark. Results show that with our R2C framework, LLMs can effectively act as low-level planners, generalizing across diverse settings and open-vocabulary robotic tasks. The code and demonstrations are available at: https://vipl-vsu.github.io/Room2Chessboard.

****

ICE: Intrinsic Concept Extraction from a Single Image via Diffusion Models

Fernando Julio Cendra, Kai Han; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23734-23743

The inherent ambiguity in defining visual concepts poses significant challenges for modern generative models, such as the diffusion-based Text-to-Image (T2I) models, in accurately learning concepts from a single image. Existing methods lack a systematic way to reliably extract the interpretable underlying intrinsic concepts. To address this challenge, we present ICE, short for Intrinsic Concept Extraction, a novel framework that exclusively utilizes a T2I model to automatically and systematically extract intrinsic concepts from a single image. ICE consists of two pivotal stages. In the first stage, ICE devises an automatic concept localization module to pinpoint relevant text-based concepts and their corresponding masks within the image. This critical stage streamlines concept initialization and provides precise guidance for subsequent analysis. The second stage delves deeper into each identified mask, decomposing the object-level concepts into intrinsic concepts and general concepts. This decomposition allows for a more granular and interpretable breakdown of visual elements. Our framework demonstrates superior performance on intrinsic concept extraction from a single image in an unsupervised manner. Project page: https://visual-ai.github.io/ice
****

ASIGN: An Anatomy-aware Spatial Imputation Graphic Network for 3D Spatial Transcriptomics

Junchao Zhu, Ruining Deng, Tianyuan Yao, Juming Xiong, Chongyu Qu, Junlin Guo, Siqi Lu, Mengmeng Yin, Yu Wang, Shilin Zhao, Haichun Yang, Yuankai Huo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30829-30838

Spatial transcriptomics (ST) is an emerging technology that enables medical computer vision scientists to automatically interpret the molecular profiles underlying morphological features. Currently, however, most deep learning-based ST analyses are limited to two-dimensional (2D) sections, which can introduce diagnostic errors due to the heterogeneity of pathological tissues across 3D sections. Expanding ST to three-dimensional (3D) volumes is challenging due to the prohibitive costs; a 2D ST acquisition already costs over 50 times more than whole slide imaging (WSI), and a full 3D volume with 10 sections can be an order of magnitude more expensive. To reduce costs, scientists have attempted to predict ST data directly from WSI without performing actual ST acquisition. However, these methods typically yield unsatisfying results. To address this, we introduce a novel problem setting: 3D ST imputation using 3D WSI histology sections combined with a single 2D ST slide. To do so, we present the Anatomy-aware Spatial Imputation Graph Network (ASIGN) for more precise, yet affordable, 3D ST modeling. The ASIGN architecture extends existing 2D spatial relationships into 3D by leveraging cross-layer overlap and similarity-based expansion. Moreover, a multi-level spatial attention graph network integrates features comprehensively across different data sources. We evaluated ASIGN on three public spatial transcriptomics datasets, with experimental results demonstrating that ASIGN achieves state-of-the-art performance on both 2D and 3D scenarios.
****

Reversing Flow for Image Restoration

Haina Qin, Wenyang Luo, Libin Wang, Dandan Zheng, Jingdong Chen, Ming Yang, Bing Li, Weiming Hu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7545-7558

Image restoration aims to recover high-quality (HQ) images from degraded low-quality (LQ) ones by reversing the effects of degradation. Existing generative models for image restoration, including diffusion and score-based models, often treat the degradation process as a stochastic transformation, which introduces inefficiency and complexity. In this work, we propose ResFlow, a novel image restoration framework that models the degradation process as a deterministic path using continuous normalizing flows. ResFlow augments the degradation process with an auxiliary process that disambiguates the uncertainty in HQ prediction to enable reversible modeling of the degradation process. ResFlow adopts entropy-preserving flow paths and learns the augmented degradation flow by matching the velocity field. ResFlow significantly improves the performance and speed of image restoration, completing the task in fewer than four sampling steps. Extensive experiment

s demonstrate that ResFlow achieves state-of-the-art results across various imag e restoration benchmarks, offering a practical and efficient solution for real-w orld applications.
****

Shadow Generation Using Diffusion Model with Geometry Prior
Haonan Zhao, Qingyang Liu, Xinhao Tao, Li Niu, Guangtao Zhai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7603-7612
Image composition involves integrating foreground object into background image t o obtain a composite image. One of the key challenges is to produce realistic sh adow for the inserted foreground object. Recently, diffusion-based methods have shown superior performance compared to GAN-based methods in shadow generation. H owever, they are still struggling to generate shadows with plausible geometry in complex cases. In this paper, we focus on promoting diffusion-based methods by leveraging geometry priors. Specifically, we first predict the rotated bounding box and matched shadow shapes for the foreground shadow. Then, the geometry info rmation of rotated bounding box and matched shadow shapes is injected into Contr olNet to facilitate shadow generation. Extensive experiments on both DESOBAv2 da taset and real composite images validate the effectiveness of our proposed metho d.
****

Rethinking Epistemic and Aleatoric Uncertainty for Active Open-Set Annotation: A n Energy-Based Approach
Chen-Chen Zong, Sheng-Jun Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10153-10162
Active learning (AL), which iteratively queries the most informative examples fr om a large pool of unlabeled candidates for model training, faces significant ch allenges in the presence of open-set classes. Existing methods either prioritize query examples likely to belong to known classes, indicating low epistemic unce rtainty (EU), or focus on querying those with highly uncertain predictions, refl ecting high aleatoric uncertainty (AU). However, they both yield suboptimal perf ormance, as low EU corresponds to limited useful information, and closed-set AU metrics for unknown class examples are less meaningful. In this paper, we propos e an Energy-based Active Open-set Annotation (EAOA) framework, which effectively integrates EU and AU to achieve superior performance. EAOA features a (C+1)-cla ss detector and a target classifier, incorporating an energy-based EU measure an d a margin-based energy loss designed for the detector, alongside an energy-base d AU measure for the target classifier. Another crucial component is the target- driven adaptive sampling strategy. It first forms a smaller candidate set with l ow EU scores to ensure closed-set properties, making AU metrics meaningful. Subs equently, examples with high AU scores are queried to form the final query set, with the candidate set size adjusted adaptively. Extensive experiments show that EAOA achieves state-of-the-art performance while maintaining high query precisi on and low training overhead. The code is available at https://github.com/chench enzong/EAOA.
****

Any3DIS: Class-Agnostic 3D Instance Segmentation by 2D Mask Tracking
Phuc Nguyen, Minh Luu, Anh Tran, Cuong Pham, Khoi Nguyen; Proceedings of the Com puter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3636-3645
Existing 3D instance segmentation methods frequently encounter issues with over- segmentation, leading to redundant and inaccurate 3D proposals that complicate d ownstream tasks. This challenge arises from their unsupervised merging approach, where dense 2D instance masks are lifted across frames into point clouds to for m 3D candidate proposals without direct supervision. These candidates are then h ierarchically merged based on heuristic criteria, often resulting in numerous re dundant segments that fail to combine into precise 3D proposals. To overcome the se limitations, we propose a 3D-Aware 2D Mask Tracking module that uses robust 3 D priors from a 2D mask segmentation and tracking foundation model (SAM-2) to en sure consistent object masks across video frames. Rather than merging all visibl e superpoints across views to create a 3D mask, our 3D Mask Optimization module leverages a dynamic programming algorithm to select an optimal set of views, ref

ining the superpoints to produce a final 3D proposal for each object. Our approa ch achieves comprehensive object coverage within the scene, significantly reduce s unnecessary proposals by half, and lowers the average runtime by a factor of 1 0, mitigating potential negative impacts on downstream applications. Evaluations on ScanNet200 and ScanNet++ confirm the effectiveness of our method, with impro vements across Class-Agnostic, Open-Vocabulary, and Open-Ended 3D Instance Segme ntation tasks.

****

FDS: Frequency-Aware Denoising Score for Text-Guided Latent Diffusion Image Edit ing

Yufan Ren, Zicong Jiang, Tong Zhang, Søren Forchhammer, Sabine Süsstrunk; Procee dings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp . 2651-2660

Text-guided image editing using Text-to-Image (T2I) models often fails to yield satisfactory results, frequently introducing unintended modifications, such as t he loss of local detail and color changes. In this paper, we analyze these failu re cases and attribute them to the indiscriminate optimization across all freque ncy bands, even though only specific frequencies may require adjustment. To addr ess this, we introduce a simple yet effective approach that enables the selectiv e optimization of specific frequency bands within localized spatial regions for precise edits. Our method leverages wavelets to decompose images into different spatial resolutions across multiple frequency bands, enabling precise modificati ons at various levels of detail. To extend the applicability of our approach, we provide a comparative analysis of different frequency-domain techniques. Additi onally, we extend our method to 3D texture editing by performing frequency decom position on the triplane representation, enabling frequency-aware adjustments fo r 3D textures. Quantitative evaluations and user studies demonstrate the effecti veness of our method in producing high-quality and precise edits.

****

MMAR: Towards Lossless Multi-Modal Auto-Regressive Probabilistic Modeling

Jian Yang, Dacheng Yin, Yizhou Zhou, Fengyun Rao, Wei Zhai, Yang Cao, Zheng-Jun Zha; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7974-7985

Recent advancements in multi-modal large language models have propelled the deve lopment of joint probabilistic models capable of both image understanding and ge neration. However, we have identified that recent methods suffer from loss of im age information during understanding task, due to either image discretization or diffusion denoising steps. To address this issue, we propose a novel Multi-Moda l Auto-Regressive (MMAR) probabilistic modeling framework. Unlike discretization line of method, MMAR takes in continuous-valued image tokens to avoid informati on loss in an efficient way. Differing from diffusion-based approaches, we disen tangle the diffusion process from auto-regressive backbone model by employing a light-weight diffusion head on top each auto-regressed image patch embedding. In this way, when the model transits from image generation to understanding throug h text generation, the backbone model's hidden representation of the image is no t limited to the last denoising step. To successfully train our method, we also propose a theoretically proven technique that addresses the numerical stability issue and a training strategy that balances the generation and understanding tas k goals. Extensive evaluations on 18 image understanding benchmarks show that MM AR significantly outperforms most of the existing joint multi-modal models, surp assing the method that employs pre-trained CLIP vision encoder. Meanwhile, MMAR is able to generate high quality images. We also show that our method is scalabl e with larger data and model size.

****

ROS-SAM: High-Quality Interactive Segmentation for Remote Sensing Moving Object

Zhe Shan, Yang Liu, Lei Zhou, Cheng Yan, Heng Wang, Xia Xie; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3625-3635

The availability of large-scale remote sensing video data underscores the import ance of high-quality interactive segmentation. However, challenges such as small object sizes, ambiguous features, and limited generalization make it difficult

for current methods to achieve this goal. In this work, we propose ROS-SAM, a me
thod designed to achieve high-quality interactive segmentation while preserving
generalization across diverse remote sensing data. The ROS-SAM is built upon thr
ee key innovations: 1) LoRA-based fine-tuning, which enables efficient domain ad
aptation while maintaining SAM's generalization ability, 2) Enhancement of netwo
rk deep layers to improve the discriminability of extracted features, thereby re
ducing misclassifications, and 3) Integration of global context with local bound
ary details in the mask decoder to generate high-quality segmentation masks. Add
itionally, we redesign the data pipeline to ensure the model learns to better ha
ndle objects at varying scales during training while focusing on high-quality pr
edictions during inference. Experiments on remote sensing video datasets show th
at the data pipeline boosts the IoU by 6%, while ROS-SAM increases the IoU by 13
%. Finally, when evaluated on existing remote sensing object tracking datasets,
ROS-SAM demonstrates impressive zero-shot capabilities, generating masks that cl
osely resemble manual annotations. These results confirm ROS-SAM as a powerful t
ool for fine-grained segmentation in remote sensing applications. Code is availa
ble at: https://github.com/ShanZard/ROS-SAM.
****

MultiMorph: On-demand Atlas Construction
S. Mazdak Abulnaga, Andrew Hoopes, Neel Dey, Malte Hoffmann, Bruce Fischl, John
Guttag, Adrian Dalca; Proceedings of the Computer Vision and Pattern Recognition
Conference (CVPR), 2025, pp. 30906-30917
We present MultiMorph, a fast and efficient method for constructing anatomical a
tlases on the fly. Atlases capture the canonical structure of a collection of im
ages and are essential for quantifying anatomical variability across populations
. However, current atlas construction methods often require days to weeks of com
putation, thereby discouraging rapid experimentation. As a result, many scientif
ic studies rely on suboptimal, precomputed atlases from mismatched populations,
negatively impacting downstream analyses. MultiMorph addresses these challenges
with a feedforward model that rapidly produces high-quality, population-specific
atlases in a single forward pass for any 3D brain dataset, without any fine-tun
ing or optimization. MultiMorph is based on a linear group-interaction layer tha
t aggregates and shares features within the group of input images. Further, by l
everaging auxiliary synthetic data, MultiMorph generalizes to new imaging modali
ties and population groups at test-time. Experimentally, MultiMorph outperforms
state-of-the-art optimization-based and learning-based atlas construction method
s in both small and large population settings, with a 100-fold reduction in time
. This makes MultiMorph an accessible framework for biomedical researchers witho
ut machine learning expertise, enabling rapid, high-quality atlas generation for
diverse studies.
****

MI-DETR: An Object Detection Model with Multi-time Inquiries Mechanism
Zhixiong Nan, Xianghong Li, Jifeng Dai, Tao Xiang; Proceedings of the Computer V
ision and Pattern Recognition Conference (CVPR), 2025, pp. 4703-4712
Based on analyzing the character of cascaded decoder architecture commonly adopt
ed in existing DETR-like models, this paper proposes a new decoder architecture.
The cascaded decoder architecture constrains object queries to update in the ca
scaded direction, only enabling object queries to learn relatively-limited infor
mation from image features. However, the challenges for object detection in natu
ral scenes (e.g., extremely-small, heavily-occluded, and confusingly mixed with
the background) require an object detection model to fully utilize image feature
s, which motivates us to propose a new decoder architecture with the parallel Mu
lti-time Inquiries (MI) mechanism. MI mechanism is very simple, enabling object
queries to parallelly perform multi-time inquiries to learn more comprehensive i
nformation from image features. Our MI based model, MI-DETR, outperforms all exi
sting DETR-like models on COCO benchmark under different backbones and training
epochs, achieving +2.3 AP and +0.6 AP improvements compared to the most represen
tative model DINO and SOTA model Relation-DETR under ResNet-50 backbone.
****

From Prototypes to General Distributions: An Efficient Curriculum for Masked Ima

ge Modeling
Jinhong Lin, Cheng-En Wu, Huanran Li, Jifan Zhang, Yu Hen Hu, Pedro Morgado; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20028-20038
Masked Image Modeling (MIM) has emerged as a powerful self-supervised learning paradigm for visual representation learning, enabling models to acquire rich visual representations by predicting masked portions of images from their visible regions. While this approach has shown promising results, we hypothesize that its effectiveness may be limited by optimization challenges during early training stages, where models are expected to learn complex image distributions from partial observations before developing basic visual processing capabilities. To address this limitation, we propose a prototype-driven curriculum leagrning framework that structures the learning process to progress from prototypical examples to more complex variations in the dataset. Our approach introduces a temperature-based annealing scheme that gradually expands the training distribution, enabling more stable and efficient learning trajectories. Through extensive experiments on ImageNet-1K, we demonstrate that our curriculum learning strategy significantly improves both training efficiency and representation quality while requiring substantially fewer training epochs compared to standard Masked Auto-Encoding. Our findings suggest that carefully controlling the order of training examples plays a crucial role in self-supervised visual learning, providing a practical solution to the early-stage optimization challenges in MIM.
****
Synthetic Visual Genome
Jae Sung Park, Zixian Ma, Linjie Li, Chenhao Zheng, Cheng-Yu Hsieh, Ximing Lu, Khyathi Chandu, Quan Kong, Norimasa Kobori, Ali Farhadi, Yejin Choi, Ranjay Krishna; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9073-9086
Reasoning over visual relationships--spatial, functional, interactional, social, etc.--are considered to be a fundamental component of human cognition.Yet, despite the major advances in visual comprehension in multimodal language models (MLMs), precise reasoning over relationships remains a challenge. We introduce Robin: an MLM instruction-tuned with densely annotated relationships capable of constructing high-quality dense scene graphs at scale. To train ROBIN, we curate SVG, a synthetic scene graph dataset by completing the missing relations of selected objects in existing scene graphs using a teacher MLM and a carefully designed filtering process to ensure high-quality. To generate more accurate and rich scene graphs at scale for any image, we introduce SG-EDIT: a self-distillation framework where GPT-4o further refines ROBIN's predicted scene graphs by removing unlikely relations and/or suggesting relevant ones. In total, our dataset contains 146K images and 5.6M relationships for 2.6M objects. Results show that our Robin-3B model, despite being trained on less than 3 million instances, outperforms similar-size models trained on over 300 million instances on relationship understanding benchmarks, and even surpasses larger models up to 13B parameters. Notably, it achieves state-of-the-art performance in referring expression comprehension with a score of 88.2, surpassing the previous best of 87.4. Our results suggest that training on the refined scene graph data is crucial to maintaining high performance across diverse visual reasoning tasks.
****
Difference Inversion: Interpolate and Isolate the Difference with Token Consistency for Image Analogy Generation
Hyunsoo Kim, Donghyun Kim, Suhyun Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18250-18259
How can we generate an image B' that satisfies A:A'::B:B', given the input images A,A' and B? Recent works have tackled this challenge through approaches like visual in-context learning or visual instruction. However, these methods are typically limited to specific models (InstructPix2Pix. Inpainting models) rather than general diffusion models (Stable Diffusion, SDXL). This dependency may lead to inherited biases or lower editing capabilities. In this paper, we propose Difference Inversion, a method that isolates only the difference from A and A' and ap

plies it to B to generate a plausible B'. To address model dependency, it is cru cial to structure prompts in the form of a "Full Prompt" suitable for input to s table diffusion models, rather than using an "Instruction Prompt". To this end, we accurately extract the Difference between A and A' and combine it with the pr ompt of B, enabling a plug-and-play application of the difference. To extract a precise difference, we first identify it through 1) Delta Interpolation. Additio nally, to ensure accurate training, we propose the 2) Token Consistency Loss and 3) Zero Initialization of Token Embeddings. Our extensive experiments demonstra te that Difference Inversion outperforms existing baselines both quantitatively and qualitatively, indicating its ability to generate more feasible B' in a mode l-agnostic manner.

****

Octopus: Alleviating Hallucination via Dynamic Contrastive Decoding

Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, Yanning Zhang; Pro ceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29904-29914

Large Vision-Language Models (LVLMs) have obtained impressive performance in vis ual content understanding and multi-modal reasoning. Unfortunately, these large models suffer from serious hallucination problems and tend to generate fabricate d responses. Recently, several Contrastive Decoding (CD) strategies have been pr oposed to alleviate hallucination by introducing disturbed inputs. Although grea t progress has been made, these CD strategies mostly apply a one-size-fits-all a pproach for all input conditions. In this paper, we revisit this process through extensive experiments. Related results show that hallucination causes are hybri d and each generative step faces a unique hallucination challenge. Leveraging th ese meaningful insights, we introduce a simple yet effective Octopus-like framew ork that enables the model to adaptively identify hallucination types and create a dynamic CD workflow. Our Octopus framework not only outperforms existing meth ods across four benchmarks but also demonstrates excellent deployability and exp ansibility. Code is available at https://github.com/LijunZhang01/Octopus.

****

MTADiffusion: Mask Text Alignment Diffusion Model for Object Inpainting

Jun Huang, Ting Liu, Yihang Wu, Xiaochao Qu, Luoqi Liu, Xiaolin Hu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1832 5-18334

Advancements in generative models have enabled image inpainting models to genera te content within specific regions of an image based on provided prompts and mas ks. However, existing inpainting methods often suffer from problems such as sema ntic misalignment, structural distortion, and style inconsistency. In this work, we present MTADiffusion, a Mask-Text Alignment diffusion model designed for obj ect inpainting. To enhance the semantic capabilities of the inpainting model, we introduce MTAPipeline, an automatic solution for annotating masks with detailed descriptions. Based on the MTAPipeline, we construct a new MTADataset comprisin g 5 million images and 25 million mask-text pairs. Furthermore, we propose a mul ti-task training strategy that integrates both inpainting and edge prediction ta sks to improve structural stability. To promote style consistency, we present a novel inpainting style-consistency loss using a pre-trained VGG network and the Gram matrix. Comprehensive evaluations on BrushBench and EditBench demonstrate t hat MTADiffusion achieves state-of-the-art performance compared to other methods .

****

Stop Learning it all to Mitigate Visual Hallucination, Focus on the Hallucinatio n Target.

Dokyoon Yoon, Youngsook Song, Woomyoung Park; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4200-4208

Multimodal Large Language Models (MLLMs) frequently suffer from hallucination is sues, generating information about objects that are not present in input images during vision-language tasks. These hallucinations particularly undermine model reliability in practical applications requiring accurate object identification. To address this challenge, we propose TL-DPO, a preference learning approach tha

t mitigates hallucinations by focusing on targeted areas where they occur. To im
plement this, we build a dataset containing hallucinated responses, correct resp
onses, and target information (i.e., objects present in the images and the corre
sponding chunk positions in responses affected by hallucinations). By applying a
 preference learning method restricted to these specific targets, the model can
filter out irrelevant signals and focus on correcting hallucinations. This allow
s the model to produce more factual responses by concentrating solely on relevan
t information. Experimental results demonstrate that TL-DPO effectively reduces
hallucinations across multiple vision hallucination tasks, improving the reliabi
lity and performance of MLLMs without diminishing overall performance.
****

Seeing the Abstract: Translating the Abstract Language for Vision Language Model
s

Davide Talon, Federico Girella, Ziyue Liu, Marco Cristani, Yiming Wang; Proceedi
ngs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp.
9253-9262

Natural language goes beyond dryly describing visual content. It contains rich a
bstract concepts to express feeling, creativity and properties that cannot be di
rectly perceived. Yet, current research in Vision Language Models (VLMs) has not
 shed light on abstract-oriented language.Our research breaks new ground by unco
vering its wide presence and under-estimated value, with extensive analysis.Part
icularly, we focus our investigation on the fashion domain, a highly-representat
ive field with abstract expressions. By analyzing recent large-scale multimodal
fashion datasets, we find that abstract terms have a dominant presence, rivaling
 the concrete ones, providing novel information, and being useful in the retriev
al task. However, a critical challenge emerges: current general-purpose or fashi
on-specific VLMs are pre-trained with databases that lack sufficient abstract wo
rds in their text corpora, thus hindering their ability to effectively represent
 abstract-oriented language. We propose a training-free and model-agnostic metho
d, Abstract-to-Concrete Translator (ACT), to shift abstract representations towa
rds well-represented concrete ones in the VLM latent space, using pre-trained mo
dels and existing multimodal databases.On the text-to-image retrieval task, desp
ite being training-free, ACT outperforms the fine-tuned VLMs in both same- and c
ross-dataset settings, exhibiting its effectiveness with a strong generalization
 capability. Moreover, the improvement introduced by ACT is consistent with vari
ous VLMs, making it a plug-and-play solution.
****

Spiking Transformer: Introducing Accurate Addition-Only Spiking Self-Attention f
or Transformer

Yufei Guo, Xiaode Liu, Yuanpei Chen, Weihang Peng, Yuhan Zhang, Zhe Ma; Proceedi
ngs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp.
24398-24408

Transformers have demonstrated outstanding performance across a wide range of ta
sks, owing to their self-attention mechanism, but they are highly energy-consumi
ng. Spiking Neural Networks have emerged as a promising energy-efficient alterna
tive to traditional Artificial Neural Networks, leveraging event-driven computat
ion and binary spikes for information transfer. The combination of Transformers'
 capabilities with the energy efficiency of SNNs offers a compelling opportunity
. This paper addresses the challenge of adapting the self-attention mechanism of
 Transformers to the spiking paradigm by introducing a novel approach: Accurate
Addition-Only Spiking Self-Attention (A^2OS^2A). Unlike existing methods that re
ly solely on binary spiking neurons for all components of the self-attention mec
hanism, our approach integrates binary, ReLU, and ternary spiking neurons. This
hybrid strategy significantly improves accuracy while preserving non-multiplicat
ive computations. Moreover, our method eliminates the need for softmax and scali
ng operations. Extensive experiments show that the A^2OS^2A-based Spiking Transf
ormer outperforms existing SNN-based Transformers on several datasets, even achi
eving an accuracy of 78.66% on ImageNet-1K. Our work represents a significant ad
vancement in SNN-based Transformer models, offering a more accurate and efficien
t solution for real-world applications.

****
Grounding 3D Object Affordance with Language Instructions, Visual Observations and Interactions
He Zhu, Quyu Kong, Kechun Xu, Xunlong Xia, Bing Deng, Jieping Ye, Rong Xiong, Yue Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17337-17346
Grounding 3D object affordance is a task that locates objects in 3D space where they can be manipulated, which links perception and action for embodied intelligence. For example, for an intelligent robot, it is necessary to accurately ground the affordance of an object and grasp it according to human instructions. In this paper, we introduce a novel task that grounds 3D object affordance based on language instructions, visual observations and interactions, which is inspired by cognitive science. We collect an Affordance Grounding dataset with Points, Images and Language instructions (AGPIL) to support the proposed task. In the 3D physical world, due to observation orientation, object rotation, or spatial occlusion, we can only get a partial observation of the object. So this dataset includes affordance estimations of objects from full-view, partial-view, and rotation-view perspectives. To accomplish this task, we propose LMAffordance3D, the first multi-modal, language-guided 3D affordance grounding network, which applies a vision-language model to fuse 2D and 3D spatial features with semantic features. Comprehensive experiments on AGPIL demonstrate the effectiveness and superiority of our method on this task, even in unseen experimental settings.
****
MOVIS: Enhancing Multi-Object Novel View Synthesis for Indoor Scenes
Ruijie Lu, Yixin Chen, Junfeng Ni, Baoxiong Jia, Yu Liu, Diwen Wan, Gang Zeng, Siyuan Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26767-26778
Repurposing pre-trained diffusion models has been proven to be effective for NVS. However, these methods are mostly limited to a single object; directly applying such methods to compositional multi-object scenarios yields inferior results, especially incorrect object placement and inconsistent shape and appearance under novel views. How to enhance and systematically evaluate the cross-view consistency of such models remains under-explored. To address this issue, we propose MOVIS to enhance the structural awareness of the view-conditioned diffusion model for multi-object NVS in terms of model inputs, auxiliary tasks, and training strategy. First, we inject structure-aware features, including depth and object mask, into the denoising U-Net to enhance the model's comprehension of object instances and their spatial relationships. Second, we introduce an auxiliary task requiring the model to simultaneously predict novel view object masks, further improving the model's capability in differentiating and placing objects. Finally, we conduct an in-depth analysis of the diffusion sampling process and carefully devise a structure-guided timestep sampling scheduler during training, which balances the learning of global object placement and fine-grained detail recovery. To systematically evaluate the plausibility of synthesized images, we propose to assess cross-view consistency and novel view object placement alongside existing image-level NVS metrics. Extensive experiments on challenging synthetic and realistic datasets demonstrate that our method exhibits strong generalization capabilities and produces consistent novel view synthesis, highlighting its potential to guide future 3D-aware multi-object NVS tasks. Our project page is available at https://jason-aplp.github.io/MOVIS/.
****
One-Step Event-Driven High-Speed Autofocus
Yuhan Bao, Shaohua Gao, Wenyong Li, Kaiwei Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6222-6230
High-speed autofocus in extreme scenes remains a significant challenge. Traditional methods rely on repeated sampling around the focus position, resulting in "focus hunting". Event-driven methods have advanced focusing speed and improved performance in low-light conditions; however, current approaches still require at least one lengthy round of "focus hunting", involving the collection of a complete focus stack. We introduce the Event Laplacian Product (ELP) focus detection f

unction, which combines event data with grayscale Laplacian information, redefining focus search as a detection task. This innovation enables the first one-step event-driven autofocus, cutting focusing time by up to two-thirds and reducing focusing error by 24 times on the DAVIS346 dataset and 22 times on the EVK4 dataset. Additionally, we present an autofocus pipeline tailored for event-only cameras, achieving accurate results across a range of challenging motion and lighting conditions. All datasets and code will be made publicly available.
****

Symbolic Representation for Any-to-Any Generative Tasks
Jiaqi Chen, Xiaoye Zhu, Yue Wang, Tianyang Liu, Xinhui Chen, Ying Chen, Chak Tou Leong, Yifei Ke, Joseph Liu, Yiwen Yuan, Julian McAuley, Li-jia Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27816-27826
We propose a symbolic generative task description language and a corresponding inference engine that can represent arbitrary multimodal tasks as structured symbolic flows. Unlike conventional generative models, which rely on large-scale training and implicit neural representations to learn cross-modal mappings--often with high computational costs and limited flexibility--our framework introduces an explicit symbolic representation composed of three core primitives: functions, parameters, and topological logic. Using a pre-trained language model, our inference engine maps natural language instructions directly to symbolic workflows in a training-free manner. Our framework successfully performs over 12 diverse multimodal generative tasks, demonstrating strong performance and flexibility without requiring task-specific tuning. Experiments show that our method not only matches or outperforms existing state-of-the-art unified models in content quality but also offers greater efficiency, editability, and interruptibility. We believe symbolic task representations provide a cost-effective and extensible foundation for advancing the capabilities of generative AI.
****

Protecting Your Video Content: Disrupting Automated Video-based LLM Annotations
Haitong Liu, Kuofeng Gao, Yang Bai, Jinmin Li, Jinxiao Shan, Tao Dai, Shu-Tao Xia; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24056-24065
Recently, video-based large language models (video-based LLMs) have achieved impressive performance across various video comprehension tasks. However, this rapid advancement raises significant privacy and security concerns, particularly regarding the unauthorized use of personal video data in automated annotation by video-based LLMs. These unauthorized annotated video-text pairs can then be used to improve the performance of downstream tasks, such as text-to-video generation. To safeguard personal videos from unauthorized use, we propose two series of protective video watermarks with imperceptible adversarial perturbations, named Ramblings and Mutes. Concretely, Ramblings aim to mislead video-based LLMs into generating inaccurate captions for the videos, thereby degrading the quality of video annotations through inconsistencies between video content and captions. Mutes, on the other hand, are designed to prompt video-based LLMs to produce exceptionally brief captions, lacking descriptive detail. Extensive experiments demonstrate that our video watermarking methods effectively protect video data by significantly reducing video annotation performance across various video-based LLMs, showcasing both stealthiness and robustness in protecting personal video content. Our code is available at https://github.com/ttthhl/Protecting_Your_Video_Content.
****

PanDA: Towards Panoramic Depth Anything with Unlabeled Panoramas and Mobius Spatial Augmentation
Zidong Cao, Jinjing Zhu, Weiming Zhang, Hao Ai, Haotian Bai, Hengshuang Zhao, Lin Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 982-992
Recently, Depth Anything Models (DAMs) - a type of depth foundation models - have demonstrated impressive zero-shot capabilities across diverse perspective images. Despite its success, it remains an open question regarding DAMs' performance

on panorama images that enjoy a large field-of-view (180x360) but suffer from s
pherical distortions. To address this gap, we conduct an empirical analysis to e
valuate the performance of DAMs on panoramic images and identify their limitatio
ns. For this, we undertake comprehensive experiments to assess the performance o
f DAMs from three key factors: panoramic representations, 360 camera positions f
or capturing scenarios, and spherical spatial transformations. This way, we reve
al some key findings, e.g., DAMs are sensitive to spatial transformations. We th
en propose a semi-supervised learning (SSL) framework to learn a panoramic DAM,
dubbed PanDA. Under the umbrella of SSL, PanDA first learns a teacher model by f
ine-tuning DAM through joint training on synthetic indoor and outdoor panoramic
datasets. Then, a student model is trained using large-scale unlabeled data, lev
eraging pseudo-labels generated by the teacher model. To enhance PanDA's general
ization capability, Mobius transformation-based spatial augmentation (MTSA) is p
roposed to impose consistency regularization between the predicted depth maps fr
om the original and spatially transformed ones. This subtly improves the student
model's robustness to various spatial transformations, even under severe distor
tions. Extensive experiments demonstrate that PanDA exhibits remarkable zero-sho
t capability across diverse scenes, and outperforms the data-specific panoramic
depth estimation methods on two popular real-world benchmarks.
****

Towards High-fidelity 3D Talking Avatar with Personalized Dynamic Texture
Xuanchen Li, Jianyu Wang, Yuhao Cheng, Yikun Zeng, Xingyu Ren, Wenhan Zhu, Weimi
ng Zhao, Yichao Yan; Proceedings of the Computer Vision and Pattern Recognition
Conference (CVPR), 2025, pp. 204-214
Significant progress has been made for speech-driven 3D face animation, but most
works focus on learning the motion of mesh/geometry, ignoring the impact of dyn
amic texture. In this work, we reveal that dynamic texture plays a key role in r
endering high-fidelity talking avatars, and introduce a high-resolution 4D datas
et TexTalk4D, consisting of 100 minutes of audio-synced scan-level meshes with d
etailed 8K dynamic textures from 100 subjects. Based on the dataset, we explore
the inherent correlation between motion and texture, and propose a diffusion-bas
ed framework TexTalker to simultaneously generate facial motions and dynamic tex
tures from speech. Furthermore, we propose a novel pivot-based style injection s
trategy to capture the complicity of different texture and motion styles, which
allows disentangled control. TexTalker, as the first method to generate audio-sy
nced facial motion with dynamic texture, not only outperforms the prior arts in
synthesising facial motions, but also produces realistic textures that are consi
stent with the underlying facial movements. Project page: https://xuanchenli.git
hub.io/TexTalk/.
****

Scene Splatter: Momentum 3D Scene Generation from Single Image with Video Diffus
ion Model
Shengjun Zhang, Jinzhao Li, Xin Fei, Hao Liu, Yueqi Duan; Proceedings of the Com
puter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6089-6098
In this paper, we propose Scene Splatter, a momentum-based paradigm for video di
ffusion to generate generic scenes from single image. Existing methods, which em
ploy video generation models to synthesize novel views, suffer from limited vide
o length and scene inconsistency, leading to artifacts and distortions during fu
rther reconstruction. To address this issue, we construct noisy samples from ori
ginal features as momentum to enhance video details and maintain scene consisten
cy. However, for latent features with the perception field that spans both known
and unknown regions, such latent-level momentum restricts the generative abilit
y of video diffusion in unknown regions. Therefore, we further introduce the afo
rementioned consistent video as a pixel-level momentum to a directly generated v
ideo without momentum for better recovery of unseen regions. Our cascaded moment
um enables video diffusion models to generate both high-fidelity and consistent
novel views. We further finetune the global Gaussian representations with enhanc
ed frames and render new frames for momentum update in the next step. In this ma
nner, we can iteratively recover a 3D scene, avoiding the limitation of video le
ngth. Extensive experiments demonstrate the generalization capability and superi

or performance of our method in high-fidelity and consistent scene generation.
****

JiSAM: Alleviate Labeling Burden and Corner Case Problems in Autonomous Driving via Minimal Real-World Data

Runjian Chen, Wenqi Shao, Bo Zhang, Shaoshuai Shi, Li Jiang, Ping Luo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6792-6801

Deep-learning-based autonomous driving (AD) perception introduces a promising picture for safe and environment-friendly transportation. However, the over-reliance on real labeled data in LiDAR perception limits the scale of on-road attempts. 3D real world data is notoriously time-and-energy-consuming to annotate and lacks corner cases like rare traffic participants. On the contrary, in simulators like CARLA, generating labeled LiDAR point clouds with corner cases is a piece of cake. However, introducing synthetic point clouds to improve real perception is non-trivial. This stems from two challenges: 1) sample efficiency of simulation datasets 2) simulation-to-real gaps. To overcome both challenges, we propose a plug-and-play method called JiSAM, shorthand for Jittering augmentation, domain-aware backbone and memory-based Sectorized AlignMent. In extensive experiments conducted on the famous AD dataset NuScenes, we demonstrate that, with SOTA 3D object detector, JiSAM is able to utilize the simulation data and only labels on 2.5% available real data to achieve comparable performance to models trained on all real data. Additionally, JiSAM achieves more than 15 mAPs on the objects not labeled in the real training set. We will release models and codes.
****

OSMamba: Omnidirectional Spectral Mamba with Dual-Domain Prior Generator for Exposure Correction

Gehui Li, Bin Chen, Chen Zhao, Lei Zhang, Jian Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7480-7490

Exposure correction is a fundamental problem in computer vision and image processing. Recently, frequency domain-based methods have achieved impressive improvement, yet they still struggle with complex real-world scenarios under extreme exposure conditions. This is due to the local convolutional receptive fields failing to model long-range dependencies in the spectrum, and the non-generative learning paradigm being inadequate for retrieving lost details from severely degraded regions. In this paper, we propose Omnidirectional Spectral Mamba (OSMamba), a novel exposure correction network that incorporates the advantages of state space models and generative diffusion models to address these limitations. Specifically, OSMamba introduces an omnidirectional spectral scanning mechanism that adapts Mamba to the frequency domain to capture comprehensive long-range dependencies in both the amplitude and phase spectra of deep image features, hence enhancing illumination correction and structure recovery. Furthermore, we develop a dual-domain prior generator that learns from well-exposed images to generate a degradation-free diffusion prior containing correct information about severely under- and over-exposed regions for better detail restoration. Extensive experiments on multiple-exposure and mixed-exposure datasets demonstrate that the proposed OSMamba achieves state-of-the-art performance both quantitatively and qualitatively.
****

Image is All You Need to Empower Large-scale Diffusion Models for In-Domain Generation

Pu Cao, Feng Zhou, Lu Yang, Tianrui Huang, Qing Song; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18358-18368

In-domain generation aims to perform a variety of tasks within a specific domain, such as unconditional generation, text-to-image, image editing, 3D generation, and more. Early research typically required training specialized generators for each unique task and domain, often relying on fully-labeled data. Motivated by the powerful generative capabilities and broad applications of diffusion models, we are driven to explore leveraging label-free data to empower these models for in-domain generation.Fine-tuning a pre-trained generative model on domain data is an intuitive but challenging way and often requires complex manual hyper-para

meter adjustments since the limited diversity of the training data can easily di srupt the model's original generative capabilities.To address this challenge, we propose a guidance-decoupled prior preservation mechanism to achieve high gener ative quality and controllability by image-only data, inspired by preserving the pre-trained model from a denoising guidance perspective.We decouple domain-rela ted guidance from the conditional guidance used in classifier-free guidance mech anisms to preserve open-world control guidance and unconditional guidance from t he pre-trained model. We further propose an efficient domain knowledge learning technique to train an additional text-free UNet copy to predict domain guidance. Besides, we theoretically illustrate a multi-guidance in-domain generation pipel ine for a variety of generative tasks, leveraging multiple guidances from distin ct diffusion models and conditions. Extensive experiments demonstrate the superi ority of our method in domain-specific synthesis and its compatibility with vari ous diffusion-based control methods and applications.

****

MedUnifier: Unifying Vision-and-Language Pre-training on Medical Data with Visio n Generation Task using Discrete Visual Representations

Ziyang Zhang, Yang Yu, Yucheng Chen, Xulei Yang, Si Yong Yeo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29744-2975 5

Despite significant progress in Vision-Language Pre-training (VLP), current appr oaches predominantly emphasize feature extraction and cross-modal comprehension, with limited attention to generating or transforming visual content. This gap h inders the model's ability to synthesize coherent and novel visual representatio ns from textual prompts, thereby reducing the effectiveness of multi-modal learn ing. In this work, we propose MedUnifier, a unified VLP framework tailored for m edical data. MedUnifier seamlessly integrates text-grounded image generation cap abilities with multi-modal learning strategies, including image-text contrastive alignment, image-text matching and image-grounded text generation. Unlike tradi tional methods that reply on continuous visual representations, our approach emp loys visual vector quantization, which not only facilitates a more cohesive lear ning strategy for cross-modal understanding but also enhances multi-modal genera tion quality by effectively leveraging discrete representations. Our framework's effectiveness is evidenced by the experiments on established benchmarks, includ ing uni-modal tasks, cross-modal tasks, and multi-modal tasks, where it achieves state-of-the-art performance across various tasks. MedUnifier also offers a hig hly adaptable tool for a wide range of language and vision tasks in healthcare, marking advancement toward the development of a generalizable AI model for medic al applications.

****

RandAR: Decoder-only Autoregressive Visual Generation in Random Orders

Ziqi Pang, Tianyuan Zhang, Fujun Luan, Yunze Man, Hao Tan, Kai Zhang, William T. Freeman, Yu-Xiong Wang; Proceedings of the Computer Vision and Pattern Recognit ion Conference (CVPR), 2025, pp. 45-55

We introduce RandAR, a decoder-only visual autoregressive (AR) model capable of generatng images in arbitrary token orders. Unlike previous decoder-only AR mode ls that rely on a predefined generation order, RandAR removes this inductive bia s, unlocking new capabilities in decoder-only generation. Our essential design e nabling random order is to insert a "position instruction token" before each ima ge token to be predicted, representing the spatial location of the next image to ken. Trained on randomly permuted token sequences -- a more challenging task tha n fixed-order generation, RandAR achieves comparable performance to conventional raster-order counterpart. More importantly, decoder-only transformers trained f rom random orders acquire new capabilities. For the efficiency bottleneck of AR models, RandAR adopts parallel decoding with KV-Cache at inference time, enjoyin g 2.5x acceleration without sacrificing generation quality. Additionally, RandAR supports in-painting, outpainting and resolution extrapolation in a zero-shot m anner.We hope RandAR inspires new directions for decoder-only visual generation models and broadens their applications across diverse scenarios. Our project pag e is at https://rand-ar.github.io/.

****
Evolving High-Quality Rendering and Reconstruction in a Unified Framework with Contribution-Adaptive Regularization
You Shen, Zhipeng Zhang, Xinyang Li, Yansong Qu, Yu Lin, Shengchuan Zhang, Liujuan Cao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16346-16355
Representing 3D scenes from multiview images is a core challenge in computer vision and graphics, which requires both precise rendering and accurate reconstruction. Recently, 3D Gaussian Splatting (3DGS) has garnered significant attention for its high-quality rendering and fast inference speed. Yet, due to the unstructured and irregular nature of Gaussian point clouds, ensuring accurate geometry reconstruction remains difficult. Existing methods primarily focus on geometry regularization, with common approaches including primitive-based and dual-model frameworks. However, the former suffers from inherent conflicts between rendering and reconstruction, while the latter is computationally and storage-intensive. To address these challenges, we propose CarGS, a unified model leveraging Contribution adaptive regularization to achieve simultaneous, high-quality rendering and surface reconstruction. The essence of our framework is learning adaptive contribution for Gaussian primitives by squeezing the knowledge from geometry regularization into a compact MLP. Additionally, we introduce a geometry-guided densification strategy with clues from both normals and Signed Distance Fields (SDF) to improve the capability of capturing high-frequency details. Our design improves the mutual learning of the two tasks, meanwhile its unified structure doesn't require separate models as in dual-model based approaches, guaranteeing efficiency. Extensive experiments demonstrate CarGS's ability to achieve state-of-the-art (SOTA) results in both rendering fidelity and reconstruction accuracy while maintaining real-time speed and minimal storage size.
****
ArticulatedGS: Self-supervised Digital Twin Modeling of Articulated Objects using 3D Gaussian Splatting
Junfu Guo, Yu Xin, Gaoyi Liu, Kai Xu, Ligang Liu, Ruizhen Hu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27144-27153
We tackle the challenge of concurrent reconstruction at the part level with the RGB appearance and estimation of motion parameters for building digital twins of articulated objects using the 3D Gaussian Splatting (3D-GS) method. With two distinct sets of multi-view imagery, each depicting an object in separate static articulation configurations, we reconstruct the articulated object in 3D Gaussian representations with both appearance and geometry information at the same time. Our approach decoupled multiple highly interdependent parameters through a multi-step optimization process, thereby achieving a stable optimization procedure and high-quality outcomes. We introduce ArticulatedGS, a self-supervised, comprehensive framework that autonomously learns to model shapes and appearances at the part level and synchronizes the optimization of motion parameters, all without reliance on 3D supervision, motion cues, or semantic labels. Our experimental results demonstrate that, among comparable methodologies, our approach has achieved optimal outcomes in terms of part segmentation accuracy, motion estimation accuracy, and visual quality.
****
NoiseCtrl: A Sampling-Algorithm-Agnostic Conditional Generation Method for Diffusion Models
Longquan Dai, He Wang, Jinhui Tang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18093-18102
In training-free conditional generative tasks, diffusion models utilize differentiable loss functions to steer the generative reverse process, necessitating modifications to sampling algorithms like DDPM and DDIM. However, such adjustments likely reduce flexibility and reliability. In this paper, we propose NoiseCtrl, a sampling-algorithm-agnostic technique for controlled image generation. Essentially, diffusion models generate denoised results z_{t-1} by adding a predicted mean μ_t with random noise ϵ_t. NoiseCtrl specifically adjusts

the random noise while leaving the underlying sampling algorithms unchanged. At each step t , NoiseCtrl converts the unconditional Gaussian noise into conditional noise epsilon'_ t by substituting the isotropic Gaussian distribution with the von Mises-Fisher distribution. This substitution introduces a directional focus while preserving the randomness required for conditional image generation. Thanks to this non-intrusive design, NoiseCtrl is straightforward to integrate and has been extensively validated through experiments, demonstrating its adaptability for different diffusion algorithms and superior performance across various conditional generation tasks.
****

Leveraging 3D Geometric Priors in 2D Rotation Symmetry Detection
Ahyun Seo, Minsu Cho; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22109-22118
Symmetry plays a vital role in understanding structural patterns, aiding object recognition and scene interpretation. This paper focuses on rotation symmetry, where objects remain unchanged when rotated around a central axis, requiring detection of rotation centers and supporting vertices. Traditional methods relied on hand-crafted feature matching, while recent segmentation models based on convolutional neural networks (CNNs) detect rotation centers but struggle with 3D geometric consistency due to viewpoint distortions. To overcome this, we propose a model that directly predicts rotation centers and vertices in 3D space and projects the results back to 2D while preserving structural integrity. By incorporating a vertex reconstruction stage enforcing 3D geometric priors--such as equal side lengths and interior angles--our model enhances robustness and accuracy. Experiments on the DENDI dataset show superior performance in rotation axis detection and validate the impact of 3D priors through ablation studies.
****

KMD: Koopman Multi-modality Decomposition for Generalized Brain Tumor Segmentation under Incomplete Modalities
Tianyi Liu, Haochuan Jiang, Kaizhu Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15663-15671
Magnetic resonance imaging (MRI), with modalities including T1, T2, T1ce, and Flair, providing complementary information critical for sub-region analysis, is widely used for brain tumor diagnosis. However, clinical practice often suffers from varying degrees of incompleteness of necessary modalities due to reasons such as susceptibility to artifacts. It significantly impairs segmentation model performance. Given the limited available modalities at hand, existing approaches attempt to project them into a shared latent space. However, they ignore decomposing the modality-shared and modality-specific information and failed to construct the relationship among different modalities. Such deficiency limits the effectiveness of the segmentation performance, particularly at the time when the amount of data in each modality is different. In this paper, we propose the plug-and-play Koopman Multi-modality Decomposition (KMD) module, leveraging the Koopman Invariant Subspace to disentangle modality-common and modality-specific information. It is capable of constructing modality relationships that minimize bias toward modalities across various modality-incomplete scenarios. More importantly, it can be integrated into several existing backbones feasibility. Through theoretical deductions and extensive empirical experiences on the BraTS2018 and BraTS2020 datasets, we have sufficiently demonstrated the effectiveness of the proposed KMD to promote generalization performance.
****

Vid2Sim: Realistic and Interactive Simulation from Video for Urban Navigation
Ziyang Xie, Zhizheng Liu, Zhenghao Peng, Wayne Wu, Bolei Zhou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1581-1591
Sim-to-real gap has long posed a significant challenge for robot learning in simulation, preventing the deployment of learned models in the real world. Previous work has primarily focused on domain randomization and system identification to mitigate this gap. However, these methods are often limited by the realism of the simulation and graphics engines. In this work, we propose Vid2Sim, a novel framework that effectively bridges the sim2real gap through a scalable and cost-ef

ficient real2sim pipeline for neural 3D scene reconstruction and simulation. Given a monocular video as input, Vid2Sim can generate photo-realistic and physically interactable 3D simulation environments to enable the reinforcement learning of visual navigation agents in complex urban environments. Extensive experiments demonstrate that Vid2Sim significantly improves the performance of urban navigation in the digital twins and real world by 31.2% and 68.3% in success rate compared with agents trained with prior simulation methods. Code and data will be made publicly available.
****

DORNet: A Degradation Oriented and Regularized Network for Blind Depth Super-Resolution

Zhengxue Wang, Zhiqiang Yan, Jinshan Pan, Guangwei Gao, Kai Zhang, Jian Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15813-15822

Recent RGB-guided depth super-resolution methods have achieved impressive performance under the assumption of fixed and known degradation (e.g., bicubic downsampling). However, in real-world scenarios, captured depth data often suffer from unconventional and unknown degradation due to sensor limitations and complex imaging environments (e.g., low reflective surfaces, varying illumination). Consequently, the performance of these methods significantly declines when real-world degradation deviate from their assumptions. In this paper, we propose the Degradation Oriented and Regularized Network (DORNet), a novel framework designed to adaptively address unknown degradation in real-world scenes through implicit degradation representations. Our approach begins with the development of a self-supervised degradation learning strategy, which models the degradation representations of low-resolution depth data using routing selection-based degradation regularization. To facilitate effective RGB-D fusion, we further introduce a degradation-oriented feature transformation module that selectively propagates RGB content into the depth data based on the learned degradation priors. Extensive experimental results on both real and synthetic datasets demonstrate the superiority of our DORNet in handling unknown degradations, outperforming existing methods.
****

Fractal Calibration for Long-tailed Object Detection

Konstantinos Panagiotis Alexandridis, Ismail Elezi, Jiankang Deng, Anh Nguyen, Shan Luo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15139-15150

Real-world datasets follow an imbalanced distribution, which poses significant challenges in rare-category object detection. Recent studies tackle this problem by developing re-weighting and re-sampling methods, that utilise the class frequencies of the dataset. However, these techniques focus solely on the frequency statistics and ignore the distribution of the classes in image space, missing important information. In contrast to them, we propose Fractal CALibration (FRACAL): a novel post-calibration method for long-tailed object detection. FRACAL devises a logit adjustment method that utilises the fractal dimension to estimate how uniformly classes are distributed in image space. During inference, it uses the fractal dimension to inversely downweight the probabilities of uniformly spaced class predictions achieving balance in two axes: between frequent and rare categories, and between uniformly spaced and sparsely spaced classes. FRACAL is a post-processing method and it does not require any training, also it can be combined with many off-the-shelf models such as one-stage sigmoid detectors and two-stage instance segmentation models. FRACAL boosts the rare class performance by up to 8.6% and surpasses all previous methods on LVIS dataset, while showing good generalisation to other datasets such as COCO, V3Det and OpenImages. We provide the code in the Appendix.
****

M3GYM: A Large-Scale Multimodal Multi-view Multi-person Pose Dataset for Fitness Activity Understanding in Real-world Settings

Qingzheng Xu, Ru Cao, Xin Shen, Heming Du, Sen Wang, Xin Yu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12289-12300

Human pose estimation is a critical task in computer vision for applications in

sports analysis, healthcare monitoring, and human-computer interaction. However, existing human pose datasets are collected either from custom-configured laboratories with complex devices or they only include data on single individuals, and both types typically capture daily activities. In this paper, we introduce the M3GYM dataset, a large-scale multimodal, multi-view, and multi-person pose dataset collected from a real gym to address the limitations of existing datasets.Specifically, we collect videos for 82 sessions from the gym, each session lasting between 40 to 60 minutes. These videos are gathered by 8 cameras, including over 50 subjects and 47 million frames. These sessions include 51 Normal fitness exercise sessions as well as 17 Pilates and 14 Yoga sessions. The exercises cover a wide range of poses and typical fitness activities, particularly in Yoga and Pilates, featuring poses with stretches, bends, and twists, e.g., humble warrior, fire hydrants and knee hover side twists. Each session involves multiple subjects, leading to significant self-occlusion and mutual occlusion in single views.Moreover, the gym has two symmetric floor mirrors, a feature not seen in previous datasets, and seven lighting conditions. We provide frame-level multimodal annotations, including 2D&3D keypoints, subject IDs, and meshes. Additionally, M3GYM uniquely offers labels for over 500 actions along with corresponding assessments from sports experts.We benchmark a variety of state-of-the-art methods for several tasks, i.e., 2D human pose estimation, single-view and multi-view 3D human pose estimation, and human mesh recovery. To simulate real-world applications, we also conduct cross-domain experiments across Normal, Yoga, and Pilates sessions. The results show that M3GYM significantly improves model generalization in complex real-world settings.

****

Noise Calibration and Spatial-Frequency Interactive Network for STEM Image Enhancement

Hesong Li, Ziqi Wu, Ruiwen Shao, Tao Zhang, Ying Fu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21287-21296

Scanning Transmission Electron Microscopy (STEM) enables the observation of atomic arrangements at sub-angstrom resolution, allowing for atomically resolved analysis of the physical and chemical properties of materials. However, due to the effects of noise, electron beam damage, sample thickness, etc, obtaining satisfactory atomic-level images is often challenging. Enhancing STEM images can reveal clearer structural details of materials. Nonetheless, existing STEM image enhancement methods usually overlook unique features in the frequency domain, and existing datasets lack realism and generality. To resolve these issues, in this paper, we develop noise calibration, data synthesis, and enhancement methods for STEM images. We first present a STEM noise calibration method, which is used to synthesize more realistic STEM images. The parameters of background noise, scan noise, and pointwise noise are obtained by statistical analysis and fitting of real STEM images containing atoms. Then we use these parameters to develop a more general dataset that considers both regular and random atomic arrangements and includes both HAADF and BF mode images. Finally, we design a spatial-frequency interactive network for STEM image enhancement, which can explore the information in the frequency domain formed by the periodicity of atomic arrangement. Experimental results show that our data is closer to real STEM images and achieves better enhancement performances together with our network. Code will be available at https://github.com/HeasonLee/SFIN.

****

Type-R: Automatically Retouching Typos for Text-to-Image Generation

Wataru Shimoda, Naoto Inoue, Daichi Haraguchi, Hayato Mitani, Seiichi Uchida, Kota Yamaguchi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2745-2754

While recent text-to-image models can generate photorealistic images from text prompts that reflect detailed instructions, they still face significant challenges in accurately rendering words in the image.In this paper, we propose to retouch erroneous text renderings in the post-processing pipeline.Our approach, called Type-R, identifies typographical errors in the generated image, erases the erroneous text, regenerates text boxes for missing words, and finally corrects typos

in the rendered words.Through extensive experiments, we show that Type-R, in combination with the latest text-to-image models such as Stable Diffusion or Flux, achieves the highest text rendering accuracy while maintaining image quality and also outperforms text-focused generation baselines in terms of balancing text accuracy and image quality.

****

Video-3D LLM: Learning Position-Aware Video Representation for 3D Scene Understanding

Duo Zheng, Shijia Huang, Liwei Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8995-9006

The rapid advancement of Multimodal Large Language Models (MLLMs) has significantly impacted various multimodal tasks. However, these models face challenges in tasks that require spatial understanding within 3D environments. Efforts to enhance MLLMs, such as incorporating point cloud features, have been made, yet a considerable gap remains between the models' learned representations and the inherent complexity of 3D scenes. This discrepancy largely stems from the training of MLLMs on predominantly 2D data, which restricts their effectiveness in comprehending 3D spaces. To address this issue, in this paper, we propose a novel generalist model, i.e., Video-3D LLM, for 3D scene understanding. By treating 3D scenes as dynamic videos and incorporating 3D position encoding into these representations, our Video-3D LLM aligns video representations with real-world spatial contexts more accurately. In addition, we have implemented a maximum coverage sampling technique to optimize the trade-off between computational cost and performance. Extensive experiments demonstrate that our model achieves state-of-the-art performance on several 3D scene understanding benchmarks, including ScanRefer, Multi3DRefer, Scan2Cap, ScanQA, and SQA3D.

****

FedSPA: Generalizable Federated Graph Learning under Homophily Heterogeneity

Zihan Tan, Guancheng Wan, Wenke Huang, He Li, Guibin Zhang, Carl Yang, Mang Ye; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15464-15475

Federated Graph Learning (FGL) has emerged as a solution to address real-world privacy concerns and data silos in graph learning, which relies on Graph Neural Networks (GNNs). Nevertheless, the homophily level discrepancies within the local graph data of clients, termed homophily heterogeneity, significantly degrade the generalizability of a global GNN. Existing research ignores this issue and suffers from unpromising collaboration. In this paper, we propose FedSPA, an effective framework that addresses homophily heterogeneity from the perspectives of homophily conflict and homophily bias. In the first place, the homophily conflict arises when training on inconsistent homophily levels across clients. Correspondingly, we propose Subgraph Feature Propagation Decoupling (SFPD), thereby achieving collaboration on unified homophily levels across clients. To further address homophily bias, we design Homophily Bias-Driven Aggregation (HBDA) which emphasizes clients with lower biases. It enables the adaptive adjustment of each client contribution to the global GNN based on its homophily bias. The superiority of FedSPA is validated through extensive experiments. The code is available at https://github.com/OakleyTan/FedSPA.

****

Homogeneous Dynamics Space for Heterogeneous Humans

Xinpeng Liu, Junxuan Liang, Chenshuo Zhang, Zixuan Cai, Cewu Lu, Yong-Lu Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27782-27793

Analyses of human motion kinematics have achieved tremendous advances. However, the production mechanism, known as human dynamics, is still undercovered. In this paper, we aim to push data-driven human dynamics understanding forward. We identify a major obstacle to this as the heterogeneity of existing human motion understanding efforts. Specifically, heterogeneity exists in not only the diverse kinematics representations and hierarchical dynamics representations but also in the data from different domains, namely biomechanics and reinforcement learning. With an in-depth analysis of the existing heterogeneity, we propose to emphasiz

e the beneath homogeneity: all of them represent the homogeneous fact of human m
otion, though from different perspectives. Given this, we propose Homogeneous Dy
namics Space (HDyS) as a fundamental space for human dynamics by aggregating het
erogeneous data and training a homogeneous latent space with inspiration from th
e inverse-forward dynamics procedure. Leveraging the heterogeneous representatio
ns and datasets, HDyS achieves decent mapping between human kinematics and dynam
ics. We demonstrate the feasibility of HDyS with extensive experiments and appli
cations. The project page is https://foruck.github.io/HDyS.
****

TailedCore: Few-Shot Sampling for Unsupervised Long-Tail Noisy Anomaly Detection
Yoon Gyo Jung, Jaewoo Park, Jaeho Yoon, Kuan-Chuan Peng, Wonchul Kim, Andrew Ben
g Jin Teoh, Octavia Camps; Proceedings of the Computer Vision and Pattern Recogn
ition Conference (CVPR), 2025, pp. 25539-25548
We aim to solve unsupervised anomaly detection in a practical challenging enviro
nment where the normal dataset is both contaminated with defective regions and i
ts product class distribution is tailed but unknown. We observe that existing mo
dels suffer from tail-versus-noise trade-off where if a model is robust against
pixel noise, then its performance deteriorates on tail class samples, and vice v
ersa. To mitigate the issue, we handle the tail class and noise samples independ
ently. To this end, we propose TailSampler, a novel class size predictor that es
timates the class cardinality of samples based on a symmetric assumption on the
class-wise distribution of embedding similarities. TailSampler can be utilized t
o sample the tail class samples exclusively, allowing to handle them separately.
 Based on these facets, we build a memory-based anomaly detection model TailedCo
re, whose memory both well captures tail class information and is noise-robust.
We extensively validate the effectiveness of TailedCore on the unsupervised long
-tail noisy anomaly detection setting, and show that TailedCore outperforms the
state-of-the-art in most settings. Code is available in TailedCore
****

GazeGene: Large-scale Synthetic Gaze Dataset with 3D Eyeball Annotations
Yiwei Bao, Zhiming Wang, Feng Lu; Proceedings of the Computer Vision and Pattern
 Recognition Conference (CVPR), 2025, pp. 18749-18759
Thanks to the introduction of large-scale datasets, deep-learning has become the
 mainstream approach for appearance-based gaze estimation problems. However, cur
rent large-scale datasets contain annotation errors and provide only a single ve
ctor for gaze annotation, lacking key information such as 3D eyeball structures.
 Limitations in annotation accuracy and variety have constrained the progress in
 research and development of deep-learning methods for appearance-based gaze-rel
ated tasks. In this paper, we present GazeGene, a new large-scale synthetic gaze
 dataset with photo-realistic samples. More importantly, GazeGene not only provi
des accurate gaze annotations, but also offers 3D annotations of vital eye struc
tures such as the pupil, iris, eyeball, optical and visual axes for the first ti
me. Experiments show that GazeGene achieves comparable quality and generalizatio
n ability with real-world datasets, even outperforms most existing datasets on h
igh-resolution images. Furthermore, its 3D eyeball annotations expand the applic
ation of deep-learning methods on various gaze-related tasks, offering new insig
hts into this field. The dataset is available at: https://phiai.buaa.edu.cn/Gaze
Gene/.
****

VideoHandles: Editing 3D Object Compositions in Videos Using Video Generative Pr
iors
Juil Koo, Paul Guerrero, Chun-Hao P. Huang, Duygu Ceylan, Minhyuk Sung; Proceedi
ngs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp.
17692-17701
Generative methods for image and video editing use generative models as priors t
o perform edits despite incomplete information, such as changing the composition
 of 3D objects shown in a single image. Recent methods have shown promising comp
osition editing results in the image setting, but in the video setting, editing
methods have focused on editing object appearance, object motion, or camera moti
on, and as a result, methods to edit object composition in videos are still miss

ing. We propose VideoHandles as a method for editing 3D object compositions in v
ideos of static scenes with camera motion. Our approach allows editing the 3D po
sition of a 3D object across all frames of a video in a temporally consistent ma
nner. This is achieved by lifting intermediate features of a generative model to
 a 3D reconstruction that is shared between all frames, editing the reconstructi
on, and projecting the features on the edited reconstruction back to each frame.
 To the best of our knowledge, this is the first generative approach to edit obj
ect compositions in videos. Our approach is simple and training-free, while outp
erforming state-of-the-art image editing baselines.
****

Satellite Observations Guided Diffusion Model for Accurate Meteorological States
 at Arbitrary Resolution

Siwei Tu, Ben Fei, Weidong Yang, Fenghua Ling, Hao Chen, Zili Liu, Kun Chen, Han
g Fan, Wanli Ouyang, Lei Bai; Proceedings of the Computer Vision and Pattern Rec
ognition Conference (CVPR), 2025, pp. 28071-28080

Accurate acquisition of surface meteorological conditions at arbitrary locations
 holds significant importance for weather forecasting and climate simulation. Me
teorological states derived from satellite observations are often provided in th
e form of low-resolution grid fields. If spatial interpolation is applied direct
ly to obtain meteorological states for specific locations, there will often be s
ignificant discrepancies compared to actual observations. Existing downscaling m
ethods for acquiring meteorological state information at higher resolutions comm
only overlook the correlation with satellite observations. To bridge the gap, we
 propose Satellite-observations Guided Diffusion Model (SGD), a conditional diff
usion model pre-trained on ERA5 reanalysis data with satellite observations (Gri
dSat) as conditions, which is employed for sampling downscaled meteorological st
ates through a zero-shot guided sampling strategy and patch-based methods. Durin
g the training process, we propose to fuse the information from GridSat satellit
e observations into ERA5 maps via the attention mechanism, enabling SGD to gener
ate atmospheric states that align more accurately with actual conditions. In the
 sampling, we employed optimizable convolutional kernels to simulate the upscale
 process, thereby generating high-resolution ERA5 maps using low-resolution ERA5
 maps as well as observations from weather stations as guidance. Moreover, our d
evised patch-based method promotes SGD to generate meteorological states at arbi
trary resolutions. Experiments demonstrate SGD fulfills accurate meteorological
states downscaling to 6.25km.
****

Reconstructing People, Places, and Cameras

Lea Müller, Hongsuk Choi, Anthony Zhang, Brent Yi, Jitendra Malik, Angjoo Kanaza
wa; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)
, 2025, pp. 21948-21958

We present "Humans and Structure from Motion" (HSfM), a method for jointly recon
structing multiple human meshes, scene point clouds, and camera parameters in a
metric world coordinate system from a sparse set of uncalibrated multi-view imag
es featuring people. Our approach combines data-driven scene reconstruction with
 the traditional Structure-from-Motion (SfM) framework to achieve more accurate
scene reconstruction and camera estimation while simultaneously recovering human
 meshes. In contrast to existing scene reconstruction and SfM methods that lack
metric scale information, our method estimates approximate metric scale by lever
aging the human statistical model. Furthermore, our method reconstructs multiple
 human meshes within the same world coordinate system with the scene point cloud
, effectively capturing spatial relationships among individuals and their positi
ons in the environment. We initialize the reconstruction of humans, scenes, and
cameras using robust foundational models and jointly optimize these elements. Th
is joint optimization synergistically improves the accuracy of each component. W
e compare our method with existing methods on two challenging benchmarks, EgoHum
ans and EgoExo4D, demonstrating significant improvements in human localization a
ccuracy within the world coordinate frame (reducing error from 3.59m to 1.04m in
 EgoHumans and from 3.01m to 0.50m in EgoExo4D). Notably, our results show that
incorporating human data into the SfM pipeline improves camera pose estimation (

e.g., increasing RRA@15 by 20.3% on EgoHumans). Additionally, qualitative results show that our approach improves scene reconstruction quality. Our code is available at muelea.github.io/hsfm.
****

InPO: Inversion Preference Optimization with Reparametrized DDIM for Efficient Diffusion Model Alignment
Yunhong Lu, Qichao Wang, Hengyuan Cao, Xierui Wang, Xiaoyin Xu, Min Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28629-28639
Without using explicit reward, direct preference optimization (DPO) employs paired human preference data to fine-tune generative models, a method that has garnered considerable attention in large language models (LLMs). However, exploration of aligning text-to-image (T2I) diffusion models with human preferences remains limited. In comparison to supervised fine-tuning, existing methods that align diffusion model suffer from low training efficiency and subpar generation quality due to the long Markov chain process and the intractability of the reverse process. To address these limitations, we introduce DDIM-InPO, an efficient method for direct preference alignment of diffusion models. Our approach conceptualizes diffusion model as a single-step generative model, allowing us to fine-tune the outputs of specific latent variables selectively. In order to accomplish this objective, we first assign implicit rewards to any latent variable directly via a reparameterization technique. Then we construct an Inversion technique to estimate appropriate latent variables for preference optimization. This modification process enables the diffusion model to only fine-tune the outputs of latent variables that have a strong correlation with the preference dataset. Experimental results indicate that our DDIM-InPO achieves state-of-the-art performance with just 400 steps of fine-tuning, surpassing all preference aligning baselines for T2I diffusion models in human preference evaluation tasks.
****

GaussTR: Foundation Model-Aligned Gaussian Transformer for Self-Supervised 3D Spatial Understanding
Haoyi Jiang, Liu Liu, Tianheng Cheng, Xinjie Wang, Tianwei Lin, Zhizhong Su, Wenyu Liu, Xinggang Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11960-11970
3D Semantic Occupancy Prediction is fundamental for spatial understanding, yet existing approaches face challenges in scalability and generalization due to their reliance on extensive labeled data and computationally intensive voxel-wise representations. In this paper, we introduce GaussTR, a novel Gaussian-based Transformer framework that unifies sparse 3D modeling with foundation model alignment through Gaussian representations to advance 3D spatial understanding. GaussTR predicts sparse sets of Gaussians in a feed-forward manner to represent 3D scenes. By splatting the Gaussians into 2D views and aligning the rendered features with foundation models, GaussTR facilitates self-supervised 3D representation learning and enables open-vocabulary semantic occupancy prediction without requiring explicit annotations. Empirical experiments on the Occ3D-nuScenes dataset demonstrate GaussTR's state-of-the-art zero-shot performance of 12.27 mIoU, along with a 40% reduction in training time. These results highlight the efficacy of GaussTR for scalable and holistic 3D spatial understanding, with promising implications in autonomous driving and embodied agents. The code is available at https://github.com/hustvl/GaussTR.
****

Single Domain Generalization for Few-Shot Counting via Universal Representation Matching
Xianing Chen, Si Huo, Borui Jiang, Hailin Hu, Xinghao Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4639-4649
Few-shot counting estimates the number of target objects in an image using only a few annotated exemplars. However, domain shift severely hinders existing methods to generalize to unseen scenarios. This falls into the realm of single domain generalization that remains unexplored in few-shot counting. To solve this problem, we begin by analyzing the main limitations of current methods, which typic

ally follow a standard pipeline that extract the object prototypes from exemplar s and then match them with image feature to construct the correlation map. We ar gue that existing methods overlook the significance of learning highly generaliz ed prototypes. Building on this insight, we propose the first domain generalizat ion few-shot counter, Universal Representation Matching, termed URM. Our primary contribution is the discovery that incorporating universal vision-language repr esentations distilled from a large scale pretrained vision-language model into t he correlation construction process substantially improves robustness to domain shifts without compromising in domain performance. As a result, URM achieves sta te-of-the-art performance on both in domain and the newly introduced domain gene ralization setting.
****

Discovering Hidden Visual Concepts Beyond Linguistic Input in Infant Learning
Xueyi Ke, Satoshi Tsutsui, Yayun Zhang, Bihan Wen; Proceedings of the Computer V ision and Pattern Recognition Conference (CVPR), 2025, pp. 4343-4352
Infants develop complex visual understanding rapidly, even preceding of the acqu isition of linguistic skills. As computer vision seeks to replicate the human vi sion system, understanding infant visual development may offer valuable insights . In this paper, we present an interdisciplinary study exploring this question: can a computational model that imitates the infant learning process develop broa der visual concepts that extend beyond the vocabulary it has heard, similar to h ow infants naturally learn? To investigate this, we analyze a recently published model in Science by Vong et al., which is trained on longitudinal, egocentric i mages of a single child paired with transcribed parental speech. We perform neur on labeling to identify visual concept neurons hidden in the model's internal re presentations. We then demonstrate that these neurons can recognize objects beyo nd the model's original vocabulary. Furthermore, we compare the differences in r epresentation between infant models and those in modern computer vision models, such as CLIP and ImageNet pre-trained model. Ultimately, our work bridges cognit ive science and computer vision by analyzing the internal representations of a c omputational model trained on an infant visual and linguistic inputs. Our code i s available at https://github.com/Kexueyi/discover_infant_vis.
****

Do We Always Need the Simplicity Bias? Looking for Optimal Inductive Biases in t he Wild
Damien Teney, Liangze Jiang, Florin Gogianu, Ehsan Abbasnejad; Proceedings of th e Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 79-90
Common choices of architecture give neural networks a preference for fitting dat a with simple functions. This simplicity bias is known as key to their success. This paper explores the limits of this assumption. Building on recent work that showed that activation functions are the origin of the simplicity bias (Teney, 2 024), we introduce a method to meta-learn activation functions to modulate this bias.**Findings.** We discover multiple tasks where the assumption of simplicity is inadequate, and standard ReLU architectures are therefore suboptimal. In the se cases, we find activation functions that perform better by inducing a prior o f higher complexity. Interestingly, these cases correspond to domains where neur al networks have historically struggled: tabular data, regression tasks, cases o f shortcut learning, and algorithmic grokking tasks. In comparison, the simplici ty bias proves adequate on image tasks, where learned activations are nearly ide ntical to ReLUs and GeLUs.**Implications.** (1) Contrary to common belief, the s implicity bias is not universally useful. There exist real tasks where it is sub optimal. (2) The suitability of ReLU models for image classification is not acci dental. (3) The success of ML ultimately depends on the adequacy between data an d architectures, and there may be benefits for architectures tailored to specifi c distributions of tasks.
****

A General Adaptive Dual-level Weighting Mechanism for Remote Sensing Pansharpeni ng
Jie Huang, Haorui Chen, Jiaxuan Ren, Siran Peng, Liangjian Deng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7447-74

56
Currently, deep learning-based methods for remote sensing pansharpening have advanced rapidly. However, many existing methods struggle to fully leverage feature heterogeneity and redundancy, thereby limiting their effectiveness. To address these challenges across two key dimensions, we introduce a general adaptive dual-level weighting mechanism (ADWM), designed to enhance a wide range of existing deep-learning methods. First, Intra-Feature Weighting (IFW) evaluates correlations among channels within each feature and selectively weighs to reduce redundancy and enhance unique information. Second, Cross-Feature Weighting (CFW) adjusts contributions across layers based on inter-layer correlations, refining the final output by preserving key distinctions across feature depths. This dual-level weighting is efficiently implemented through our proposed Correlation-Aware Covariance Weighting (CACW), which generates weights by utilizing the correlations captured within the covariance matrix. Extensive experiments demonstrate the superior performance of ADWM compared to recent state-of-the-art (SOTA) methods. Furthermore, we validate the effectiveness of our approach through generality experiments, ablation studies, comparison experiments, and detailed visual analysis.
****

RASP: Revisiting 3D Anamorphic Art for Shadow-Guided Packing of Irregular Objects
Soumyaratna Debnath, Ashish Tiwari, Kaustubh Sadekar, Shanmuganathan Raman; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5849-5858
Recent advancements in learning-based methods have opened new avenues for exploring and interpreting art forms, such as shadow art, origami, and sketch art, through computational models. One notable visual art form is 3D Anamorphic Art in which an ensemble of arbitrarily shaped 3D objects creates a realistic and meaningful expression when observed from a particular viewpoint and loses its coherence over the other viewpoints. In this work, we build on insights from 3D Anamorphic Art to perform 3D object arrangement. We introduce RASP, a differentiable-rendering-based framework to arrange arbitrarily shaped 3D objects within a bounded volume via shadow (or silhouette)-guided optimization with an aim of minimal inter-object spacing and near-maximal occupancy. Furthermore, we propose a novel SDF-based formulation to handle inter-object intersection and container extrusion. We demonstrate that RASP can be extended to part assembly alongside object packing considering 3D objects to be "parts" of another 3D object. Finally, we present artistic illustrations of multi-view anamorphic art, achieving meaningful expressions from multiple viewpoints within a single ensemble.
****

Identifying and Mitigating Spurious Correlation in Multi-Task Learning
Junyi Chai, Shenyu Lu, Xiaoqian Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25698-25707
Multi-task learning (MTL) is a paradigm that aims to improve the generalization of models by simultaneously learning multiple related tasks, leveraging shared representations and task-specific information to enhance performance on individual tasks. However, existing work has shown that MTL can potentially hinder generalization, with one key factor being spurious correlations between tasks. Owing to the knowledge-sharing property, the per-task predictors are more likely to develop reliance on spurious features. Most existing approaches address this issue through distributional robustness, aiming to maintain consistent performance across different distributions under unknown covariate shifts. However, this formulation lacks theoretical guarantees and can be sensitive to the construction of covariate shifts. In this work, we propose a novel perspective, where we seek to identify spurious correlations between tasks. Drawing inspirations from conventional formulations on spurious correlation, for each task, we propose to distinguish its spurious tasks using the difference in correlation coefficients between the empirical distribution and class-wise resampled distributions, thereby capturing the correlations between task labels w.r.t. each class. We prove theoretically the feasibility of the resampling strategy in characterizing spurious correlations between tasks. Furthermore, we propose a simple fine-tuning strategy, deb

iased adversarial training, where the per-task predictors are adversarially trai
ned to disregard information associated with their spurious tasks. Experimental
results on six benchmark datasets show that our method effectively mitigates spu
rious correlations and outperforms state-of-the-art methods in improving general
ization.
****

Continuous, Subject-Specific Attribute Control in T2I Models by Identifying Sema
ntic Directions
Stefan Andreas Baumann, Felix Krause, Michael Neumayr, Nick Stracke, Melvin Sevi
, Vincent Tao Hu, Björn Ommer; Proceedings of the Computer Vision and Pattern Re
cognition Conference (CVPR), 2025, pp. 13231-13241
Recent advances in text-to-image (T2I) diffusion models have significantly impro
ved the quality of generated images. However, providing efficient control over i
ndividual subjects, particularly the attributes characterizing them, remains a k
ey challenge. While existing methods have introduced mechanisms to modulate attr
ibute expression, they typically provide either detailed, object-specific locali
zation of such a modification or full-scale fine-grained, nuanced control of att
ributes. No current approach offers both simultaneously, resulting in a gap when
 trying to achieve precise continuous and subject-specific attribute modulation
in image generation. In this work, we demonstrate that token-level directions ex
ist within commonly used CLIP text embeddings that enable fine-grained, subject-
specific control of high-level attributes in T2I models. We introduce two method
s to identify these directions: a simple, optimization-free technique and a lear
ning-based approach that utilizes the T2I model to characterize semantic concept
s more specifically. Our methods allow the augmentation of the prompt text input
, enabling fine-grained control over multiple attributes of individual subjects
simultaneously, without requiring any modifications to the diffusion model itsel
f. This approach offers a unified solution that fills the gap between global and
 localized control, providing competitive flexibility and precision in text-guid
ed image generation.
****

Diffusion Bridge: Leveraging Diffusion Model to Reduce the Modality Gap Between
Text and Vision for Zero-Shot Image Captioning
Jeong Ryong Lee, Yejee Shin, Geonhui Son, Dosik Hwang; Proceedings of the Comput
er Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4050-4059
The modality gap between vision and text embeddings in CLIP presents a significa
nt challenge for zero-shot image captioning, limiting effective cross-modal repr
esentation. Traditional approaches, such as noise injection and memory-based sim
ilarity matching, attempt to address this gap, yet these methods either rely on
indirect alignment or relatively naive solutions with heavy computation. Diffusi
on Bridge introduces a novel approach to directly reduce this modality gap by le
veraging Denoising Diffusion Probabilistic Models (DDPM), trained exclusively on
 text embeddings to model their distribution. Our approach is motivated by the o
bservation that, while paired vision and text embeddings are relatively close, a
 modality gap still exists due to stable regions created by the contrastive loss
. This gap can be interpreted as noise in cross-modal mappings, which we approxi
mate as Gaussian noise. To bridge this gap, we employ a reverse diffusion proces
s, where image embeddings are strategically introduced at an intermediate step i
n the reverse process, allowing them to be refined progressively toward the text
 embedding distribution. This process transforms vision embeddings into text-lik
e representations closely aligned with paired text embeddings, effectively minim
izing discrepancies between modalities. Experimental results demonstrate that th
ese text-like vision embeddings significantly enhance alignment with their paire
d text embeddings, leading to improved zero-shot captioning performance on MSCOC
O and Flickr30K. Diffusion Bridge achieves competitive results without reliance
on memory banks or entity-driven methods, offering a novel pathway for cross-mod
al alignment and opening new possibilities for the application of diffusion mode
ls in multi-modal tasks. The source code is available at: https://github.com/mon
georoo/diffusion-bridge
****

MODfinity: Unsupervised Domain Adaptation with Multimodal Information Flow Intertwining
Shanglin Liu, Jianming Lv, Jingdan Kang, Huaidong Zhang, Zequan Liang, Shengfeng He; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5092-5101
Multimodal unsupervised domain adaptation leverages unlabeled data in the target domain to enhance multimodal systems continuously. While current state-of-the-art methods encourage interaction between sub-models of different modalities through pseudo-labeling and feature-level exchange, varying sample quality across modalities can lead to the propagation of inaccurate information, resulting in error accumulation. To address this, we propose Modal-Affinity Multimodal Domain Adaptation (MODfinity), a method that dynamically manages multimodal information flow through fine-grained control over teacher model selection, guiding information intertwining at both feature and label levels. By treating labels as an independent modality, MODfinity enables balanced performance assessment across modalities, employing a novel modal-affinity measurement to evaluate information quality. Additionally, we introduce a modal-affinity distillation technique to control sample-level information exchange, ensuring reliable multimodal interaction based on affinity evaluations within the feature space. Extensive experiments on three multimodal datasets demonstrate that our framework consistently outperforms state-of-the-art methods, particularly in high-noise environments.
****
Towards Universal Soccer Video Understanding
Jiayuan Rao, Haoning Wu, Hao Jiang, Ya Zhang, Yanfeng Wang, Weidi Xie; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8384-8394
As a globally celebrated sport, soccer has attracted widespread interest from fans over the world. This paper aims to develop a comprehensive multi-modal framework for soccer video understanding.Specifically, we make the following contributions in this paper:(i) we introduce **SoccerReplay-1988**, the largest multi-modal soccer dataset to date, featuring videos and detailed annotations from 1,988 complete matches, with an automated annotation pipeline;(ii) we present the first visual-language foundation model in the soccer domain, **MatchVision**, which leverages spatiotemporal information across soccer videos and excels in various downstream tasks;(iii) we conduct extensive experiments and ablation studies on action classification, commentary generation, and multi-view foul recognition,and demonstrate state-of-the-art performance on all of them, substantially outperforming existing models, which has demonstrated the superiority of our proposed data and model. We believe that this work will offer a standard paradigm for sports understanding research. The code and model will be publicly available for reproduction.
****
SimLingo: Vision-Only Closed-Loop Autonomous Driving with Language-Action Alignment
Katrin Renz, Long Chen, Elahe Arani, Oleg Sinavski; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11993-12003
Integrating large language models (LLMs) into autonomous driving has attracted significant attention with the hope of improving generalization and explainability. However, existing methods often focus on either driving or vision-language understanding but achieving both high driving performance and extensive language understanding remains challenging. In addition, the dominant approach to tackle vision-language understanding is using visual question answering. However, for autonomous driving, this is only useful if it is aligned with the action space. Otherwise, the model's answers could be inconsistent with its behavior. Therefore, we propose a model that can handle three different tasks: (1) closed-loop driving, (2) vision-language understanding, and (3) language-action alignment. Our model SimLingo is based on a vision language model (VLM) and works using only camera, excluding expensive sensors like LiDAR. SimLingo obtains state-of-the-art performance on the widely used CARLA simulator on the Bench2Drive benchmark and is the winning entry at the CARLA challenge 2024. Additionally, we achieve strong

results in a wide variety of language-related tasks while maintaining high drivi
ng performance.
****

Improved Video VAE for Latent Video Diffusion Model
Pingyu Wu, Kai Zhu, Yu Liu, Liming Zhao, Wei Zhai, Yang Cao, Zheng-Jun Zha; Proc
eedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025,
pp. 18124-18133
Variational Autoencoder (VAE) aims to compress pixel data into low-dimensional l
atent space, playing an important role in OpenAI's Sora and other latent video d
iffusion generation models. While most existing video VAEs inflate a pre-trained
 image VAE into the 3D causal structure for temporal-spatial compression, this p
aper presents two astonishing findings: (1) The initialization from a well-train
ed image VAE with the same latent dimensions is not an optimal scheme. (2) The a
doption of causal reasoning leads to unequal information interactions and unbala
nced performance between frames. To alleviate these problems, we propose a keyfr
ame-based temporal compression (KTC) architecture and a group causal convolution
 (GCConv) module to further improve video VAE (IV-VAE). Specifically, the KTC ar
chitecture divides the latent space into two branches, in which one half complet
ely inherits the compression prior of keyframes from a lower-dimension image VAE
 while the other half involves temporal compression through 3D group causal conv
olution, reducing temporal-spatial conflicts and accelerating the convergence sp
eed of video VAE. The GCConv in the above 3D half uses standard convolution with
in each frame group to ensure inter-frame equivalence, and employs causal logica
l padding between groups to maintain flexibility in processing variable frame vi
deo. Extensive experiments on five benchmarks demonstrate the SOTA video reconst
ruction and generation abilities of our IV-VAE.
****

Immune: Improving Safety Against Jailbreaks in Multi-modal LLMs via Inference-Ti
me Alignment
Soumya Suvra Ghosal, Souradip Chakraborty, Vaibhav Singh, Tianrui Guan, Mengdi W
ang, Ahmad Beirami, Furong Huang, Alvaro Velasquez, Dinesh Manocha, Amrit Singh
Bedi; Proceedings of the Computer Vision and Pattern Recognition Conference (CVP
R), 2025, pp. 25038-25049
With the widespread deployment of Multimodal Large Language Models (MLLMs) for v
isual-reasoning tasks, improving their safety has become crucial. Recent researc
h indicates that despite training-time safety alignment, these models remain vul
nerable to jailbreak attacks--carefully crafted image-prompt pairs that compel t
he model to generate harmful content. In this work, we first highlight a critica
l safety gap, demonstrating that alignment achieved solely through safety traini
ng may be insufficient against jailbreak attacks. To address this vulnerability,
 we propose Immune, an inference-time defense framework that leverages a safe re
ward model during decoding to defend against jailbreak attacks. Additionally, we
 provide a mathematical characterization of Immune, offering insights on why it
improves safety against jailbreak. Extensive evaluations on diverse jailbreak be
nchmarks using recent MLLMs reveal that Immune effectively enhances model safety
 while preserving the model's original capabilities. For instance, against text-
based jailbreak attacks on LLaVA-1.6, Immune reduces the attack success rate by
57.82% and 16.78% compared to the base MLLM and state-of-the-art defense strateg
y, respectively.
****

Efficient Video Super-Resolution for Real-time Rendering with Decoupled G-buffer
 Guidance
Mingjun Zheng, Long Sun, Jiangxin Dong, Jinshan Pan; Proceedings of the Computer
 Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11328-11337
Latency is a key driver for real-time rendering applications, making super-resol
ution techniques increasingly popular to accelerate rendering processes. In cont
rast to existing methods that directly concatenate low-resolution frames and G-b
uffers as input without discrimination, we develop an asymmetric UNet-based supe
r-resolution network with decoupled G-buffer guidance, dubbed RDG, to facilitate
 the spatial and temporal feature exploration for minimizing performance overhea

ds and latency.We first propose a dynamic feature modulator (DFM) to selectively encode the spatial information to capture precise structural information.We then incorporate auxiliary G-buffer information to guide the decoder to generate detail-rich, temporally stable results.Specifically, we adopt a high-frequency feature booster (HFB) to adaptively transfer the high-frequency information from the normal and bidirectional reflectance distribution function (BRDF) components of the G-buffer, enhancing the details of the generated results.To further enhance the temporal stability, we design a cross-frame temporal refiner (CTR) with depth and motion vector constraints to aggregate the previous and current frames.Extensive experimental results reveal that our proposed method is capable of generating high-quality and temporally stable results in real-time rendering.The proposed RDG-s produces 1080P rendering results on a RTX 3090 GPU with a speed of 126 FPS.

****

CustomKD: Customizing Large Vision Foundation for Edge Model Improvement via Knowledge Distillation

Jungsoo Lee, Debasmit Das, Munawar Hayat, Sungha Choi, Kyuwoong Hwang, Fatih Porikli; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25176-25186

We propose a novel knowledge distillation approach, CustomKD, that effectively leverages large vision foundation models (LVFMs) to enhance the performance of edge models (e.g., MobileNetV3). Despite recent advancements in LVFMs, such as DINOv2 and CLIP, their potential in knowledge distillation for enhancing edge models remains underexplored. While knowledge distillation is a promising approach for improving the performance of edge models, the discrepancy in model capacities and heterogeneous architectures between LVFMs and edge models poses a significant challenge. Our observation indicates that although utilizing larger backbones (e.g., ViT-S to ViT-L) in teacher models improves their downstream task performances, the knowledge distillation from the large teacher models fails to bring as much performance gain for student models as for teacher models due to the large model discrepancy. Our simple yet effective CustomKD customizes the well-generalized features inherent in LVFMs to a given student model in order to reduce model discrepancies. Specifically, beyond providing well-generalized original knowledge from teachers, CustomKD aligns the features of teachers to those of students, making it easy for students to understand and overcome the large model discrepancy overall. CustomKD significantly improves the performances of edge models in scenarios with unlabeled data such as unsupervised domain adaptation (e.g., OfficeHome and DomainNet) and semi-supervised learning (e.g., CIFAR-100 with 400 labeled samples and ImageNet with 1% labeled samples), achieving the new state-of-the-art performances.

****

Enhancing Privacy-Utility Trade-offs to Mitigate Memorization in Diffusion Models

Chen Chen, Daochang Liu, Mubarak Shah, Chang Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8182-8191

Text-to-image diffusion models have demonstrated remarkable capabilities in creating images highly aligned with user prompts, yet their proclivity for memorizing training set images has sparked concerns about the originality of the generated images and privacy issues, potentially leading to legal complications for both model owners and users, particularly when the memorized images contain proprietary content. Although methods to mitigate these issues have been suggested, enhancing privacy often results in a significant decrease in the utility of the outputs, as indicated by text-alignment scores. To bridge the research gap, we introduce a novel method, PRSS, which refines the classifier-free guidance approach in diffusion models by integrating prompt re-anchoring (PR) to improve privacy and incorporating semantic prompt search (SS) to enhance utility. Extensive experiments across various privacy levels demonstrate that our approach consistently improves the privacy-utility trade-off, establishing a new state-of-the-art.

****

Learned Image Compression with Dictionary-based Entropy Model

Jingbo Lu, Leheng Zhang, Xingyu Zhou, Mu Li, Wen Li, Shuhang Gu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12850-12859

Learned image compression methods have attracted great research interest and exhibited superior rate-distortion performance to the best classical image compression standards of the present.The entropy model plays a key role in learned image compression, which estimates the probability distribution of the latent representation for further entropy coding.Most existing methods employed hyper-prior and auto-regressive architectures to form their entropy models.However, they only aimed to explore the internal dependencies of latent representation while neglecting the importance of extracting prior from training data.In this work, we propose a novel entropy model named Dictionary-based Cross Attention Entropy model, which introduces a learnable dictionary to summarize the typical structures occurring in the training dataset to enhance the entropy model.Extensive experimental results have demonstrated that the proposed model strikes a better balance between performance and latency, achieving state-of-the-art results on various benchmark datasets.

****

PMNI: Pose-free Multi-view Normal Integration for Reflective and Textureless Surface Reconstruction

Mingzhi Pei, Xu Cao, Xiangyi Wang, Heng Guo, Zhanyu Ma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26834-26843

Reflective and textureless surfaces remain a challenge in multi-view 3D reconstruction. Both camera pose calibration and shape reconstruction often fail due to insufficient or unreliable cross-view visual features. To address these issues, we present PMNI (Pose-free Multi-view Normal Integration), a neural surface reconstruction method that incorporates rich geometric information by leveraging surface normal maps instead of RGB images. By enforcing geometric constraints from surface normals and multi-view shape consistency within a neural signed distance function (SDF) optimization framework, PMNI simultaneously recovers accurate camera poses and high-fidelity surface geometry. Experimental results on synthetic and real-world datasets show that our method achieves state-of-the-art performance in the reconstruction of reflective surfaces, even without reliable initial camera poses.

****

NVComposer: Boosting Generative Novel View Synthesis with Multiple Sparse and Unposed Images

Lingen Li, Zhaoyang Zhang, Yaowei Li, Jiale Xu, Wenbo Hu, Xiaoyu Li, Weihao Cheng, Jinwei Gu, Tianfan Xue, Ying Shan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 777-787

Recent advancements in generative models have significantly improved novel view synthesis (NVS) from multi-view data. However, existing methods depend on external multi-view alignment processes, such as explicit pose estimation or pre-reconstruction, which limits their flexibility and accessibility, especially when alignment is unstable due to insufficient overlap or occlusions between views. In this paper, we propose NVComposer, a novel approach that eliminates the need for explicit external alignment. NVComposer enables the generative model to implicitly infer spatial and geometric relationships between multiple conditional views by introducing two key components: 1) an image-pose dual-stream diffusion model that simultaneously generates target novel views and condition camera poses, and 2) a geometry-aware feature alignment module that distills geometric priors from dense stereo models during training. Extensive experiments demonstrate that NVComposer achieves state-of-the-art performance in generative multi-view NVS tasks, removing the reliance on external alignment and thus improving model accessibility. Our approach shows substantial improvements in synthesis quality as the number of unposed input views increases, highlighting its potential for more flexible and accessible generative NVS systems.

****

LeanGaussian: Breaking Pixel or Point Cloud Correspondence in Modeling 3D Gaussians

Jiamin Wu, Kenkun Liu, Han Gao, Xiaoke Jiang, Yuan Yao, Lei Zhang; Proceedings o f the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26641 -26651

Rencently, Gaussian splatting has demonstrated significant success in novel view synthesis. Current methods often regress Gaussians with pixel or point cloud co rrespondence, linking each Gaussian with a pixel or a 3D point. This leads to th e redundancy of Gaussians being used to overfit the correspondence rather than t he objects represented by the 3D Gaussians themselves, consequently wasting reso urces and lacking accurate geometries or textures.In this paper, we introduce Le anGaussian, a novel approach that treats each query in deformable Transformer as one 3D Gaussian ellipsoid, breaking the pixel or point cloud correspondence con straints. We leverage deformable decoder to iteratively refine the Gaussians lay er-by-layer with the image features as keys and values.Notably, the center of ea ch 3D Gaussian is defined as 3D reference points, which are then projected onto the image for deformable attention in 2D space.On both the ShapeNet SRN dataset (category level) and the Google Scanned Objects dataset (open-category level, tr ained with the Objaverse dataset), our approach, outperforms prior methods by ap proximately 6.1%, achieving a PSNR of 25.44 and 22.36, respectively. Additionall y, our method achieves a 3D reconstruction speed of 7.2 FPS and rendering speed 500 FPS.

****

Modeling Multiple Normal Action Representations for Error Detection in Procedura l Tasks

Wei-Jin Huang, Yuan-Ming Li, Zhi-Wei Xia, Yu-Ming Tang, Kun-Yu Lin, Jian-Fang Hu , Wei-Shi Zheng; Proceedings of the Computer Vision and Pattern Recognition Conf erence (CVPR), 2025, pp. 27794-27804

Error detection in procedural activities is essential for consistent and correct outcomes in AR-assisted and robotic systems. Existing methods often focus on te mporal ordering errors or rely on static prototypes to represent normal actions. However, these approaches typically overlook the common scenario where multiple , distinct actions are valid following a given sequence of executed actions. Thi s leads to two issues: (1) the model cannot effectively detect errors using stat ic prototypes when the inference environment or action execution distribution di ffers from training; and (2) the model may also use the wrong prototypes to dete ct errors if the ongoing action label is not the same as the predicted one. To a ddress this problem, we propose an Adaptive Multiple Normal Action Representatio n (AMNAR) framework. AMNAR predicts all valid next actions and reconstructs thei r corresponding normal action representations, which are compared against the on going action to detect errors. Extensive experiments demonstrate that AMNAR achi eves state-of-the-art performance, highlighting the effectiveness of AMNAR and t he importance of modeling multiple valid next actions in error detection. The co de is available at https://github.com/iSEE-Laboratory/AMNAR.

****

Efficient Personalization of Quantized Diffusion Model without Backpropagation

Hoigi Seo, Wongi Jeong, Kyungryeol Lee, Se Young Chun; Proceedings of the Comput er Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7717-7727

Diffusion models have shown remarkable performance in image synthesis, but they demand extensive computational and memory resources for training, fine-tuning an d inference. Although advanced quantization techniques have successfully minimiz ed memory usage for inference, training and fine-tuning these quantized models s till require large memory possibly due to dequantization for accurate computatio n of gradients and/or backpropagation for gradient-based algorithms. However, me mory-efficient fine-tuning is particularly desirable for applications such as pe rsonalization that often must be run on edge devices like mobile phones with pri vate data. In this work, we address this challenge by quantizing a diffusion mod el with personalization via Textual Inversion and by leveraging a zeroth-order o ptimization on personalization tokens without dequantization so that it does not require gradient and activation storage for backpropagation that consumes consi derable memory. Since a gradient estimation using zeroth-order optimization is q uite noisy for a single or a few images in personalization, we propose to denois

e the estimated gradient by projecting it onto a subspace that is constructed wi th the past history of the tokens, dubbed Subspace Gradient. In addition, we inv estigated the influence of text embedding in image generation, leading to our pr oposed time steps sampling, dubbed Partial Uniform Timestep Sampling for samplin g with effective diffusion timesteps. Our method achieves comparable performance to prior methods in image and text alignment scores for personalizing Stable Di ffusion with only forward passes while reducing training memory demand up to 8.2 x. Project page: https://ignoww.github.io/ZOODiP_project/
****

Alias-Free Latent Diffusion Models: Improving Fractional Shift Equivariance of D iffusion Latent Space
Yifan Zhou, Zeqi Xiao, Shuai Yang, Xingang Pan; Proceedings of the Computer Visi on and Pattern Recognition Conference (CVPR), 2025, pp. 34-44
Latent Diffusion Models (LDMs) are known to have an unstable generation process, where even small perturbations or shifts in the input noise can lead to signifi cantly different outputs. This hinders their applicability in applications requi ring consistent results. In this work, we redesign LDMs to enhance consistency b y making them shift-equivariant. While introducing anti-aliasing operations can partially improve shift-equivariance, significant aliasing and inconsistency per sist due to the unique challenges in LDMs, including 1) aliasing amplification d uring VAE training and multiple U-Net inferences, and 2) self-attention modules that inherently lack shift-equivariance. To address these issues, we redesign th e attention modules to be shift-equivariant and propose an equivariance loss tha t effectively suppresses the frequency bandwidth of the features in the continuo us domain. The resulting alias-free LDM (AF-LDM) achieves strong shift-equivaria nce and is also robust to irregular warping. Extensive experiments demonstrate t hat AF-LDM produces significantly more consistent results than vanilla LDM acros s various applications, including video editing and image-to-image translation.
****

A Unified Latent Schrodinger Bridge Diffusion Model for Unsupervised Anomaly Det ection and Localization
Shilhora Akshay, Niveditha Lakshmi Narasimhan, Jacob George, Vineeth N Balasubra manian; Proceedings of the Computer Vision and Pattern Recognition Conference (C VPR), 2025, pp. 25528-25538
Anomaly detection and localization remain pivotal challenges in computer vision, with applications ranging from industrial inspection to medical diagnostics. Wh ile current supervised methods offer high precision, they are often impractical due to the scarcity of annotated data and the infrequent occurrence of anomalies . Recent advancements in unsupervised approaches, particularly reconstruction-ba sed methods, have addressed these issues by training models exclusively on norma l data, enabling them to identify anomalies during inference. However, these met hods frequently rely on auxiliary networks or specialized adaptations, which can limit their robustness and practicality. This work introduces the Latent Anomal y Schrodinger Bridge (LASB), a unified unsupervised anomaly detection model that operates entirely in the latent space without requiring additional networks or custom modifications. LASB transforms anomaly images into normal images by prese rving structural integrity across varying anomaly classes, lighting, and pose co nditions, making it highly robust and versatile. Unlike previous methods, LASB d oes not focus solely on reconstructing anomaly features, but emphasizes anomaly transformation, achieving smooth anomaly-to-normal image conversions. Our method achieves state-of-the-art performance on both the MVTec-AD and VisA datasets, e xcelling in detection and localization tasks.
****

KVQ: Boosting Video Quality Assessment via Saliency-guided Local Perception
Yunpeng Qu, Kun Yuan, Qizhi Xie, Ming Sun, Chao Zhou, Jian Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2150-21 60
Video Quality Assessment (VQA), which intends to predict the perceptual quality of videos, has attracted increasing attention. Due to factors like motion blur o r specific distortions, the quality of different regions in a video varies. Reco

gnizing the region-wise local quality within a video is beneficial for assessing global quality and can guide us in adopting fine-grained enhancement or transcoding strategies. Due to the heavy cost of annotating region-wise quality, the lack of ground truth constraints from relevant datasets further complicates the utilization of local perception. Inspired by the Human Visual System (HVS) that links global quality to the local texture of different regions and their visual saliency, we propose a Kaleidoscope Video Quality Assessment (KVQ) framework, which aims to effectively assess both saliency and local texture, thereby facilitating the assessment of global quality. Our framework extracts visual saliency and allocates attention using Fusion-Window Attention (FWA) while incorporating a Local Perception Constraint (LPC) to mitigate the reliance of regional texture perception on neighboring areas. KVQ obtains significant improvements across multiple scenarios on five VQA benchmarks compared to SOTA methods. Furthermore, to assess local perception, we establish a new Local Perception Visual Quality (LPVQ) dataset with region-wise annotations. Experimental results demonstrate the capability of KVQ in perceiving local distortions. KVQ models and the LPVQ dataset will be available at https://github.com/qyp2000/KVQ.

****

MambaVision: A Hybrid Mamba-Transformer Vision Backbone

Ali Hatamizadeh, Jan Kautz; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25261-25270

We propose a novel hybrid Mamba-Transformer backbone, MambaVision, specifically tailored for vision applications. Our core contribution includes redesigning the Mamba formulation to enhance its capability for efficient modeling of visual features. Through a comprehensive ablation study, we demonstrate the feasibility of integrating Vision Transformers (ViT) with Mamba. Our results show that equipping the Mamba architecture with self-attention blocks in the final layers greatly improves its capacity to capture long-range spatial dependencies. Based on these findings, we introduce a family of MambaVision models with a hierarchical architecture to meet various design criteria. For classification on the ImageNet-1K dataset, MambaVision variants achieve state-of-the-art (SOTA) performance in terms of both Top-1 accuracy and throughput. In downstream tasks such as object detection, instance segmentation, and semantic segmentation on MS COCO and ADE20K datasets, MambaVision outperforms comparably sized backbones while demonstrating favorable performance. Code: https://github.com/NVlabs/MambaVision

****

Learning Flow Fields in Attention for Controllable Person Image Generation

Zijian Zhou, Shikun Liu, Xiao Han, Haozhe Liu, Kam Woh Ng, Tian Xie, Yuren Cong, Hang Li, Mengmeng Xu, Juan-Manuel Perez-Rua, Aditya Patel, Tao Xiang, Miaojing Shi, Sen He; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2491-2501

Controllable person image generation aims to generate a person image conditioned on reference images, allowing precise control over the person's appearance or pose.However, prior methods often distort fine-grained textural details from the reference image, despite achieving high overall image quality.We attribute these distortions to inadequate attention to corresponding regions in the reference image.To address this, we thereby propose learning flow fields in attention (Leffa), which explicitly guides the target query to attend to the correct reference key in the attention layer during training.Specifically, it is realized via a regularization loss on top of the attention map within a diffusion-based baseline. Our extensive experiments show that Leffa achieves state-of-the-art performance in controlling appearance (virtual try-on) and pose (pose transfer), significantly reducing fine-grained detail distortion while maintaining high image quality. Additionally, we show that our loss is model-agnostic and can be used to improve the performance of other diffusion models.

****

Multi-Label Prototype Visual Spatial Search for Weakly Supervised Semantic Segmentation

Songsong Duan, Xi Yang, Nannan Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30241-30250

Existing Weakly Supervised Semantic Segmentation (WSSS) relies on the CNN-based Class Activation Map (CAM) and Transformer-based self-attention map to generate class-specific masks for semantic segmentation. However, CAM and self-attention maps usually cause incomplete segmentation due to classification bias issue. To address this issue, we propose a Multi-Label Prototype Visual Spatial Search (MuP-VSS) method with a spatial query mechanism, which learns a set of learnable class token vectors as queries to search the similarity visual tokens from image patch tokens. Specifically, MuP-VSS consists of two key components: multi-label prototype representation and multi-label prototype optimization. The former designs a global embedding to learn the global tokens from the images, and then proposes a Prototype Embedding Module (PEM) to interact with patch tokens to understand the local semantic information. The latter utilizes the exclusivity and consistency principles of the multi-label prototypes to design three prototype losses to optimize them, which contain cross-class prototype (CCP) contrastive loss, cross-image prototype (CIP) contrastive loss, and patch-to-prototype (P2P) consistency loss. CCP loss models exclusivity of multi-label prototypes learned from a single image to enhance the discriminative properties of each class better. CCP loss learns the consistency of the same class-specific prototypes extracted from multiple images to enhance the semantic consistency. P2P loss is proposed to control the semantic response of the prototype to the image patches. Experimental results on Pascal VOC 2012 and MS COCO show that MuP-VSS significantly outperforms recent methods and achieves state-of-the-art performance.
****

Early-Bird Diffusion: Investigating and Leveraging Timestep-Aware Early-Bird Tickets in Diffusion Models for Efficient Training

Lexington Whalen, Zhenbang Du, Haoran You, Chaojian Li, Sixu Li, Yingyan Lin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7675-7684

Training diffusion models (DMs) requires substantial computational resources due to multiple forward and backward passes across numerous timesteps, motivating research into efficient training techniques. In this paper, we propose EB-Diff-Train, a new efficient DM training approach that is orthogonal to other methods of accelerating DM training, by investigating and leveraging Early-Bird (EB) tickets--sparse subnetworks that manifest early in the training process and maintain high generation quality. We first investigate the existence of traditional EB tickets in DMs, enabling competitive generation quality without fully training a dense model. Then, we delve into the concept of diffusion-dedicated EB tickets, drawing on insights from varying importance of different timestep regions. These tickets adapt their sparsity levels according to the importance of corresponding timestep regions, allowing for aggressive sparsity during non-critical regions while conserving computational resources for crucial timestep regions. Building on this, we develop an efficient DM training technique that derives timestep-aware EB tickets, trains them in parallel, and combines them during inference for image generation. Extensive experiments validate the existence of both traditional and timestep-aware EB tickets, as well as the effectiveness of our proposed EB-Diff-Train method. This approach can significantly reduce training time both spatially and temporally--achieving 2.9x 5.8x speedups over training unpruned dense models, and up to 10.3x faster training compared to standard train-prune-finetune pipelines--without compromising generative quality. Our code is available at https://github.com/GATECH-EIC/Early-Bird-Diffusion.
****

FireEdit: Fine-grained Instruction-based Image Editing via Region-aware Vision Language Model

Jun Zhou, Jiahao Li, Zunnan Xu, Hanhui Li, Yiji Cheng, Fa-Ting Hong, Qin Lin, Qinglin Lu, Xiaodan Liang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13093-13103

Currently, instruction-based image editing methods have made significant progress by leveraging the powerful cross-modal understanding capabilities of visual language models (VLMs). However, they still face challenges in three key areas: 1) complex scenarios; 2) semantic consistency; and 3) fine-grained editing. To add

ress these issues, we propose FireEdit, an innovative Fine-grained Instruction-based image editing framework that exploits a REgion-aware VLM. FireEdit is designed to accurately comprehend user instructions and ensure effective control over the editing process. Specifically, we enhance the fine-grained visual perception capabilities of the VLM by introducing additional region tokens. Relying solely on the output of the LLM to guide the diffusion model may lead to suboptimal editing results. Therefore, we propose a Time-Aware Target Injection module and a Hybrid Visual Cross Attention module. The former dynamically adjusts the guidance strength at various denoising stages by integrating timestep embeddings with the text embeddings. The latter enhances visual details for image editing, thereby preserving semantic consistency between the edited result and the source image. By combining the VLM enhanced with fine-grained region tokens and the time-dependent diffusion model, FireEdit demonstrates significant advantages in comprehending editing instructions and maintaining high semantic consistency. Extensive experiments indicate that our approach surpasses the state-of-the-art instruction-based image editing methods.

****

Doppelgangers++: Improved Visual Disambiguation with Geometric 3D Features

Yuanbo Xiangli, Ruojin Cai, Hanyu Chen, Jeffrey Byrne, Noah Snavely; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27166-27175

Accurate 3D reconstruction is frequently hindered by visual aliasing, where visually similar but distinct surfaces (aka, doppelgangers), are incorrectly matched. These spurious matches distort the structure-from-motion (SfM) process, leading to misplaced model elements and reduced accuracy. Prior efforts addressed this with CNN classifiers trained on curated datasets, but these approaches struggle to generalize across diverse real-world scenes and can require extensive parameter tuning. In this work, we present Doppelgangers++, a method to enhance doppelganger detection and improve 3D reconstruction accuracy. Our contributions include a diversified training dataset that incorporates geo-tagged images from everyday scenes to expand robustness beyond landmark-based datasets. We further propose a Transformer-based classifier that leverages 3D-aware features from the MASt3R model, achieving superior precision and recall across both in-domain and out-of-domain tests. Doppelgangers++ integrates seamlessly into standard SfM and MASt3R-SfM pipelines, offering efficiency and adaptability across varied scenes. To evaluate SfM accuracy, we introduce an automated, geotag-based method for validating reconstructed models, eliminating the need for manual inspection. Through extensive experiments, we demonstrate that Doppelgangers++ significantly enhances pairwise visual disambiguation and improves 3D reconstruction quality in complex and diverse scenarios.

****

Learnable Infinite Taylor Gaussian for Dynamic View Rendering

Bingbing Hu, Yanyan Li, Rui Xie, Bo Xu, Haoye Dong, Junfeng Yao, Gim Hee Lee; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26844-26854

Capturing the temporal evolution of Gaussian properties such as position, rotation, and scale is a challenging task due to the vast number of time-varying parameters and the limited photometric data available, which generally results in convergence issues, making it difficult to find an optimal solution. While feeding all inputs into an end-to-end neural network can effectively model complex temporal dynamics, this approach lacks explicit supervision and struggles to generate high-quality transformation fields. On the other hand, using time-conditioned polynomial functions to model Gaussian trajectories and orientations provides a more explicit and interpretable solution, but requires significant handcrafted effort and lacks generalizability across diverse scenes. To overcome these limitations, this paper introduces a novel approach based on a learnable infinite Taylor Formula to model the temporal evolution of Gaussians. This method offers both the flexibility of an implicit network-based approach and the interpretability of explicit polynomial functions, allowing for more robust and generalizable modeling of Gaussian dynamics across various dynamic scenes.Extensive experiments on

dynamic novel view rendering tasks are conducted on public datasets, demonstrating that the proposed method achieves state-of-the-art performance in this domain.

****

DL2G: Degradation-guided Local-to-Global Restoration for Eyeglass Reflection Removal

Zhilv Yi, Xiao Lu, Hong Ding, Jingbo Hu, Zhi Jiang, Chunxia Xiao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16061-16070

Eyeglass reflection removal can restore the texture information in the reflection destructed eye area, which is meaningful for various tasks on the facial images. It is still challenging to correctly eliminate reflections, reasonably restore the lost contents, and guarantee that the final result has a consistent color and illumination with the input image. In this paper, we introduce a Degradation-guided Local-to-Global (DL2G) restoration framework to address this problem. We first propose a multiplicative reflection degradation model, which is used to alleviate reflection degradation to obtain a preliminary result. Then, in the local details restoration stage, we propose a local structure-aware diffusion model to learn the true distribution of texture details in the eye area. This helps in recovering lost contents in the regions of heavy degradation where the background is invisible. Finally, in the global consistency refinement stage, we utilize the input image as a reference image to generate the final result that is consistent with the input image in color and illumination. Extensive experiments demonstrate that our method can improve the effect of reflection removal and generate results with more reasonable semantics, exquisite details, and harmonious illumination.

****

DIV-FF: Dynamic Image-Video Feature Fields For Environment Understanding in Egocentric Videos

Lorenzo Mur-Labadia, Josechu Guerrero, Ruben Martinez-Cantin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3470-3480

Environment understanding in egocentric videos is an important step for applications like robotics, augmented reality and assistive technologies. These videos are characterized by dynamic interactions and a strong dependence on the wearer's engagement with the environment. Traditional approaches often focus on isolated clips or fail to integrate rich semantic and geometric information, limiting scene comprehension. We introduce Dynamic Image-Video Feature Fields (DIV-FF), a framework that decomposes the egocentric scene into persistent, dynamic, and actor-based components while integrating both image and video-language features. Our model enables detailed segmentation, captures affordances, understands the surroundings and maintains consistent understanding over time. DIV-FF outperforms state-of-the-art methods, particularly in dynamically evolving scenarios, demonstrating its potential to advance long-term, spatio-temporal scene understanding.

****

SaMam: Style-aware State Space Model for Arbitrary Image Style Transfer

Hongda Liu, Longguang Wang, Ye Zhang, Ziru Yu, Yulan Guo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28468-28478

Global effective receptive field plays a crucial role for image style transfer (ST) to obtain high-quality stylized results. However, existing ST backbones (e.g., CNNs and Transformers) suffer huge computational complexity to achieve global receptive fields. Recently, the State Space Model (SSM), especially the improved variant Mamba, has shown great potential for long-range dependency modeling with linear complexity, which offers a approach to resolve the above dilemma. In this paper, we develop a Mamba-based style transfer framework, termed SaMam. Specifically, a mamba encoder is designed to efficiently extract content and style information. In addition, a style-aware mamba decoder is developed to flexibly adapt to various styles. Moreover, to address the problems of local pixel forgetting, channel redundancy and spatial discontinuity of existing SSMs, we introduce both local enhancement and zigzag scan. Qualitative and quantitative results demonstrate that our SaMam outperforms state-of-the-art methods in terms of both ac

curacy and efficiency.
****
MFogHub: Bridging Multi-Regional and Multi-Satellite Data for Global Marine Fog Detection and Forecasting
Mengqiu Xu, Kaixin Chen, Heng Guo, Yixiang Huang, Ming Wu, Zhenwei Shi, Chuang Zhang, Jun Guo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12637-12646
Deep learning approaches for marine fog detection and forecasting have outperformed traditional methods, demonstrating significant scientific and practical importance. However, the limited availability of open-source datasets remains a major challenge. Existing datasets, often focused on a single region or satellite, restrict the ability to evaluate model performance across diverse conditions and hinder the exploration of intrinsic marine fog characteristics. To address these limitations, we introduce MFogHub, the first multi-regional and multi-satellite dataset to integrate annotated marine fog observations from 15 coastal fog-prone regions and six geostationary satellites, comprising over 68,000 high-resolution samples. By encompassing diverse regions and satellite perspectives, MFogHub facilitates rigorous evaluation of both detection and forecasting methods under varying conditions. Extensive experiments with 16 baseline models demonstrate that MFogHub can reveal generalization fluctuations due to regional and satellite discrepancy, while also serving as a valuable resource for the development of targeted and scalable fog prediction techniques. Through MFogHub, we aim to advance both the practical monitoring and scientific understanding of marine fog dynamics on a global scale. The dataset and code are available in the supplementary materials.
****
The Illusion of Unlearning: The Unstable Nature of Machine Unlearning in Text-to-Image Diffusion Models
Naveen George, Karthik Nandan Dasaraju, Rutheesh Reddy Chittepu, Konda Reddy Mopuri; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13393-13402
Text-to-image models such as Stable Diffusion, DALL*E, and Midjourney have gained immense popularity lately. However, they are trained on vast amounts of data that may include private, explicit, or copyrighted material used without permission, raising serious legal and ethical concerns. In light of the recent regulations aimed at protecting individual data privacy, there has been a surge in Machine Unlearning methods designed to remove specific concepts from these models. However, we identify a critical flaw in these unlearning techniques: unlearned concepts will revive when the models are fine-tuned, even with general or unrelated prompts. In this paper, for the first time, through an extensive study, we demonstrate the unstable nature of existing unlearning methods in text-to-image diffusion models. We introduce a framework that includes a couple of measures for analyzing the stability of existing unlearning methods. Further, the paper offers preliminary insights into the plausible explanation for the instability of the mapping-based unlearning methods that can guide future research toward more robust unlearning techniques. Codes for implementing the proposed framework are provided.
****
Making Old Film Great Again: Degradation-aware State Space Model for Old Film Restoration
Yudong Mao, Hao Luo, Zhiwei Zhong, Peilin Chen, Zhijiang Zhang, Shiqi Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28039-28049
Unlike modern native digital videos, the restoration of old films requires addressing specific degradations inherent to analog sources. However, existing specialized methods still fall short compared to general video restoration techniques. In this work, we propose a new baseline to re-examine the challenges in old film restoration. First, we develop an improved Mamba-based framework, dubbed MambaOFR, which can dynamically adjust the degradation removal patterns by generating degradation-aware prompts to tackle the complex and composite degradations pres

ent in old films. Second, we introduce a flow-guided mask deformable alignment module to mitigate the propagation of structured defect features in the temporal domain. Third, we introduce the first benchmark dataset that includes both synthetic and real-world old film clips. Extensive experiments show that the proposed method achieves state-of-the-art performance, outperforming existing advanced approaches in old film restoration. The implementation and model is available at https://github.com/MaoAYD/MambaOFR.

****

Leveraging Global Stereo Consistency for Category-Level Shape and 6D Pose Estimation from Stereo Images

Junning Qiu, Minglei Lu, Fei Wang, Yu Guo, Yonggen Ling; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16839-16849

Stereo-based category-level shape and 6D pose estimation methods have the potential to generalize to a wider range of materials than RGBD methods, which often suffer from depth measurement errors. However, without explicit depth from two views, parameters to be estimated can become inherently entangled, negatively impacting performance. To address this, we propose a method that leverages global stereo consistency to constrain optimization directions and mitigate parameter entanglement. We first estimate an intra-category occupancy field to represent a unified shape across views, ensuring consistency and preventing shape ambiguity. Through a divide-and-conquer approach within global shape fitting, we fit this shape to stereo images to obtain the pose, iteratively rendering normalized depth maps and exchanging information across views. This approach improves convergence toward the correct pose and scale. We validated our method on both depth-friendly and depth-challenging materials using our S-RGBD dataset and the TOD benchmark. Our method surpasses RGBD methods on challenging objects and performs comparably on depth-friendly ones. Ablation studies confirm the effectiveness of each component.

****

AlphaPre: Amplitude-Phase Disentanglement Model for Precipitation Nowcasting

Kenghong Lin, Baoquan Zhang, Demin Yu, Wenzhi Feng, Shidong Chen, Feifan Gao, Xutao Li, Yunming Ye; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17841-17850

Precipitation nowcasting involves using current radar observation sequences to predict future radar sequences and determine future precipitation distribution, which is crucial for disaster warning, traffic planning, and agricultural production. Despite numerous advancements, challenges persist in accurately predicting both the location and intensity of precipitation, as these factors are often interdependent, with complex atmospheric dynamics and moisture distribution causing position and intensity changes to be intricately coupled. Inspired by the fact that in the frequency domain, phase variations are shown to correspond to changes in the position of precipitation, while amplitude variations are linked to intensity changes, we propose an amplitude-phase disentanglement model called AlphaPre, which separately learn the position and intensity changes of precipitation. AlphaPre comprises three key components: a phase network, an amplitude network, and an AlphaMixer. The phase network captures positional changes by learning phase variations, while the amplitude network models intensity changes by alternating between the frequency and spatial domains. The AlphaMixer then integrates these components to produce a refined precipitation forecast. Extensive experiments on four datasets demonstrate the effectiveness and superiority of our method over state-of-the-art approaches. Our code is publicly available at https://github.com/linkenghong/AlphaPre.

****

EfficientLLaVA: Generalizable Auto-Pruning for Large Vision-language Models

Yinan Liang, Ziwei Wang, Xiuwei Xu, Jie Zhou, Jiwen Lu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9445-9454

While multimodal large language models demonstrate strong performance in complex reasoning tasks, they pose significant challenges related to model complexity during deployment, especially for resource-limited devices. In this paper, we propose an automatic pruning method for large vision-language models to enhance the

efficiency of multimodal reasoning. Conventional methods rely on the training data of the original model to select the proper pruning ratio for different network components. However, these methods are impractical for large vision-language models due to the unaffordable search costs caused by web-scale training corpus. In contrast, our approach only leverages a small number of samples to search for the desired pruning policy by maximizing its generalization ability on unknown training data while maintaining the model accuracy, which enables the achievement of an optimal trade-off between accuracy and efficiency for large visual language models. Specifically, we formulate the generalization gap of the pruning strategy using the structural risk minimization principle. Based on both task performance and generalization capability, we iteratively search for the optimal pruning policy within a given search space and optimize the vision projector to evolve the search space with higher upper bound of performance. We conduct extensive experiments on the ScienceQA, Vizwiz, MM-vet, and LLaVA-Bench datasets for the task of visual question answering. Using only 64 samples for pruning policy search, EfficientLLaVA achieves an accuracy of 83.05% on ScienceQA, along with a x1.8 speedup compared to the dense LLaVA-v1.5-7B model.

****

Detection-Friendly Nonuniformity Correction: A Union Framework for Infrared UAV Target Detection

Houzhang Fang, Xiaolin Wang, Zengyang Li, Lu Wang, Qingshan Li, Yi Chang, Luxin Yan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11898-11907

Infrared unmanned aerial vehicle (UAV) images captured using thermal detectors are often affected by temperature-dependent low-frequency nonuniformity, which significantly reduces the contrast of the images. Detecting UAV targets under nonuniform conditions is crucial in UAV surveillance applications. Existing methods typically treat infrared nonuniformity correction (NUC) as a preprocessing step for detection, which leads to suboptimal performance. Balancing the two tasks while enhancing detection-beneficial information remains challenging. In this paper, we present a detection-friendly union framework, termed UniCD, that simultaneously addresses both infrared NUC and UAV target detection tasks in an end-to-end manner. We first model NUC as a small number of parameter estimation problem jointly driven by priors and data to generate detection-conducive images. Then, we incorporate a new auxiliary loss with target mask supervision into the backbone of the infrared UAV target detection network to strengthen target features while suppressing the background. To better balance correction and detection, we introduce a detection-guided self-supervised loss to reduce feature discrepancies between the two tasks, thereby enhancing detection robustness to varying nonuniformity levels. Additionally, we construct a new benchmark composed of 50,000 infrared images in various nonuniformity types, multi-scale UAV targets and rich backgrounds with target annotations, called IRBFD. Extensive experiments on IRBFD demonstrate that our UniCD is a robust union framework for NUC and UAV target detection while achieving real-time processing capabilities. Dataset can be available at https://github.com/IVPLaboratory/UniCD.

****

Articulated Kinematics Distillation from Video Diffusion Models

Xuan Li, Qianli Ma, Tsung-Yi Lin, Yongxin Chen, Chenfanfu Jiang, Ming-Yu Liu, Donglai Xiang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17571-17581

We present Articulated Kinematics Distillation (AKD), a framework for generating high-fidelity character animations by merging the strengths of skeleton-based animation and modern generative models. AKD uses a skeleton-based representation for rigged 3D assets, drastically reducing the Degrees of Freedom (DoFs) by focusing on joint-level control, which allows for efficient, consistent motion synthesis. Through Score Distillation Sampling (SDS) with pre-trained video diffusion models, AKD distills complex, articulated motions while maintaining structural integrity, overcoming challenges faced by 4D neural deformation fields in preserving shape consistency. This approach is naturally compatible with physics-based simulation, ensuring physically plausible interactions. Experiments show that A

KD achieves superior 3D consistency and motion quality compared with existing works on text-to-4D generation.
****

ExpertAF: Expert Actionable Feedback from Video
Kumar Ashutosh, Tushar Nagarajan, Georgios Pavlakos, Kris Kitani, Kristen Grauman; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13582-13594
Feedback is essential for learning a new skill or improving one's current skill-level. However, current methods for skill-assessment from video only provide scores or compare demonstrations, leaving the burden of knowing what to do differently on the user. We introduce a novel method to generate actionable feedback (AF) from video of a person doing a physical activity, such as basketball or soccer. Our method takes a video demonstration and its accompanying 3D body pose and generates (1) free-form expert commentary describing what the person is doing well and what they could improve, and (2) a visual expert demonstration that incorporates the required corrections. We show how to leverage Ego-Exo4D's videos of skilled activity and expert commentary together with a strong language model to create a weakly-supervised training dataset for this task, and we devise a multimodal video-language model to infer coaching feedback. Our method is able to reason across multi-modal input combinations to output full-spectrum, actionable coaching--expert commentary, expert video retrieval, and expert pose generation--outperforming strong vision-language models on both established metrics and human preference studies.
****

MP-SfM: Monocular Surface Priors for Robust Structure-from-Motion
Zador Pataki, Paul-Edouard Sarlin, Johannes L. Schönberger, Marc Pollefeys; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21891-21901
While Structure-from-Motion (SfM) has seen much progress over the years, state-of-the-art systems are prone to failure when facing extreme viewpoint changes in low-overlap, low-parallax or high-symmetry scenarios. Because capturing images that avoid these pitfalls is challenging, this severely limits the wider use of SfM, especially by non-expert users. We overcome these limitations by augmenting the classical SfM paradigm with monocular depth and normal priors inferred by deep neural networks. Thanks to a tight integration of monocular and multi-view constraints, our approach significantly outperforms existing ones under extreme viewpoint changes, while maintaining strong performance in standard conditions. We also show that monocular priors can help reject faulty associations due to symmetries, which is a long-standing problem for SfM. This makes our approach the first capable of reliably reconstructing challenging indoor environments from few images. Through principled uncertainty propagation, it is robust to errors in the priors, can handle priors inferred by different models with little tuning, and will thus easily benefit from future progress in monocular depth and normal estimation. Our code is publicly available at https://github.com/cvg/mpsfm.
****

OnlineAnySeg: Online Zero-Shot 3D Segmentation by Visual Foundation Model Guided 2D Mask Merging
Yijie Tang, Jiazhao Zhang, Yuqing Lan, Yulan Guo, Dezun Dong, Chenyang Zhu, Kai Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3676-3685
Online 3D open-vocabulary segmentation of a progressively reconstructed scene is both a critical and challenging task for embodied applications. With the success of visual foundation models (VFMs) in the image domain, leveraging 2D priors to address 3D online segmentation has become a prominent research focus. Since segmentation results provided by 2D priors often require spatial consistency to be lifted into final 3D segmentation, an efficient method for identifying spatial overlap among 2D masks is essential--yet existing methods rarely achieve this in real time, mainly limiting its use to offline approaches. To address this, we propose an efficient method that lifts 2D masks generated by VFMs into a unified 3D instance using a hashing technique. By employing voxel hashing for efficient

3D scene querying, our approach reduces the time complexity of costly spatial overlap queries from O(n^2) to O(n). Accurate spatial associations further enable 3D merging of 2D masks through simple similarity-based filtering in a zero-shot manner, making our approach more robust to incomplete and noisy data. Evaluated on the ScanNet and SceneNN benchmarks, our approach achieves state-of-the-art performance in online, open-vocabulary 3D instance segmentation with leading efficiency.

****

Tora: Trajectory-oriented Diffusion Transformer for Video Generation

Zhenghao Zhang, Junchao Liao, Menghao Li, ZuoZhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, Weizhi Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2063-2073

Recent advancements in Diffusion Transformer (DiT) have demonstrated remarkable proficiency in producing high-quality video content. Nonetheless, the potential of transformer-based diffusion models for effectively generating videos with controllable motion remains an area of limited exploration. This paper introduces Tora, the first trajectory-oriented DiT framework that concurrently integrates textual, visual, and trajectory conditions, thereby enabling scalable video generation with effective motion guidance. Specifically, Tora consists of a Trajectory Extractor (TE), a Spatial-Temporal DiT, and a Motion-guidance Fuser (MGF). The TE encodes arbitrary trajectories into hierarchical spacetime motion patches with a 3D motion compression network. The MGF integrates the motion patches into the DiT blocks to generate consistent videos that accurately follow designated trajectories. Our design aligns seamlessly with DiT's scalability, allowing precise control of video content's dynamics with diverse durations, aspect ratios, and resolutions. Extensive experiments demonstrate that Tora excels in achieving high motion fidelity compared to the foundational DiT model, while also accurately simulating the complex movements of the physical world. Code is made available at https://github.com/alibaba/Tora .

****

Volumetrically Consistent 3D Gaussian Rasterization

Chinmay Talegaonkar, Yash Belhe, Ravi Ramamoorthi, Nicholas Antipa; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10953-10963

Recently, 3D Gaussian Splatting (3DGS) has enabled photorealistic view synthesis at high inference speeds. However, its splatting-based rendering model makes several approximations to the rendering equation, reducing physical accuracy. We show that splatting and its approximations are unnecessary, even within a rasterizer; we instead volumetrically integrate 3D Gaussians directly to compute the transmittance across them analytically. We use this analytic transmittance to derive more physically accurate alpha values than 3DGS, which can directly be used within their framework. The result is a method that more closely follows the volume rendering equation (similar to ray tracing) while enjoying the speed benefits of rasterization. Our method represents opaque surfaces with higher accuracy and fewer points than 3DGS. This enables it to outperform 3DGS for view synthesis (measured in SSIM and LPIPS). Being volumetrically consistent also enables our method to work out of the box for tomography. We match the state-of-the-art 3DGS-based tomography method with fewer points. Our code is publicly available at: https://github.com/chinmay0301ucsd/Vol3DGS

****

Deterministic-to-Stochastic Diverse Latent Feature Mapping for Human Motion Synthesis

Yu Hua, Weiming Liu, Gui Xu, Yaqing Hou, Yew-Soon Ong, Qiang Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22724-22734

Human motion synthesis aims to generate plausible human motion sequences, which has raised widespread attention in computer animation. Recent score-based generative models (SGMs) have demonstrated impressive results on this task. However, their training process involves complex curvature trajectories, leading to unstable training process.In this paper, we propose a Deterministic-to-Stochastic Dive

rse Latent Feature Mapping (DSDFM) method for human motion synthesis.DSDFM consi sts of two stages. The first human motion reconstruction stage aims to learn the latent space distribution of human motions. The second diverse motion generatio n stage aims to build connections between the Gaussian distribution and the late nt space distribution of human motions, thereby enhancing the diversity and accu racy of the generated human motions. This stage is achieved by the designed dete rministic feature mapping procedure with DerODE and stochastic diverse output ge neration procedure with DivSDE. DSDFM is easy to train compared to previous SGMs -based methods and can enhance diversity without introducing additional training parameters.Through qualitative and quantitative experiments, DSDFM achieves sta te-of-the-art results surpassing the latest methods, validating its superiority in human motion synthesis.

****

Morpheus: Text-Driven 3D Gaussian Splat Shape and Color Stylization

Jamie Wynn, Zawar Qureshi, Jakub Powierza, Jamie Watson, Mohamed Sayed; Proceedi ngs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7825-7836

Exploring real-world spaces using novel-view synthesis is fun, and reimagining t hose worlds in a different style adds another layer of excitement. Stylized worl ds can also be used for downstream tasks where there is limited training data an d a need to expand a model's training distribution. Most current novel-view synt hesis stylization techniques lack the ability to convincingly change geometry. T his is because any geometry change requires increased style strength which is of ten capped for stylization stability and consistency. In this work, we propose a new autoregressive 3D Gaussian Splatting stylization method. As part of this me thod, we contribute a new RGBD diffusion model that allows for strength control over appearance and shape stylization. To ensure consistency across stylized fra mes, we use a combination of novel depth-guided cross attention, feature injecti on, and a Warp ControlNet conditioned on composite frames for guiding the styliz ation of new frames. We validate our method via extensive qualitative results, q uantitative experiments, and a user study.

****

CacheQuant: Comprehensively Accelerated Diffusion Models

Xuewen Liu, Zhikai Li, Qingyi Gu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23269-23280

Diffusion models have gradually gained prominence in the field of image synthesi s, showcasing remarkable generative capabilities. Nevertheless, the slow inferen ce and complex networks, resulting from redundancy at both temporal and structur al levels, hinder their low-latency applications in real-world scenarios. Curren t acceleration methods for diffusion models focus separately on temporal and str uctural levels. However, independent optimization at each level to further push the acceleration limits results in significant performance degradation. On the o ther hand, integrating optimizations at both levels can compound the acceleratio n effects. Unfortunately, we find that the optimizations at these two levels are not entirely orthogonal. Performing separate optimizations and then simply inte grating them results in unsatisfactory performance. To tackle this issue, we pro pose CacheQuant, a novel training-free paradigm that comprehensively accelerates diffusion models by jointly optimizing model caching and quantization technique s. Specifically, we employ a dynamic programming approach to determine the optim al cache schedule, in which the properties of caching and quantization are caref ully considered to minimize errors. Additionally, we propose decoupled error cor rection to further mitigate the coupled and accumulated errors step by step. Exp erimental results show that CacheQuant achieves a 5.18xspeedup and 4xcompression for Stable Diffusion on MS-COCO, with only a 0.02 loss in CLIP score.

****

The Impact Label Noise and Choice of Threshold has on Cross-Entropy and Soft-Dic e in Image Segmentation

Marcus Nordström, Atsuto Maki, Henrik Hult; Proceedings of the Computer Vision a nd Pattern Recognition Conference (CVPR), 2025, pp. 20820-20829

In image segmentation and specifically in medical image segmentation, the soft-D

ice loss is often chosen instead of the more traditional cross-entropy loss to improve performance with respect to the Dice metric.Experimental work supporting this claim exists, but how and why the two loss functions lead to different predictions is not well understood.This paper explains the observed discrepancy as a consequence of how those loss functions are effected by label noise and what threshold is used to convert the predicted soft segmentation to predicted labels.In particular, it is shown (i) how the optimal solutions to the two loss functions diverge as the noise is increased, and (ii) how the optimal solutions to soft-Dice can be recovered by thresholding the solutions to cross-entropy with an a priori unknown but efficiently computable threshold.The theoretical results are supported by numerical experiments and it is concluded that cross-entropy with the alternative threshold yields the stability and informative label probability maps associated with cross-entropy without sacrificing the performance of soft-Dice.

****

Open-World Objectness Modeling Unifies Novel Object Detection

Shan Zhang, Yao Ni, Jinhao Du, Yuan Xue, Philip Torr, Piotr Koniusz, Anton van den Hengel; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30332-30342

The challenge in open-world object detection, similarly to few- and zero-shot learning, is to generalize beyond the class distribution of the training data. In this paper, we propose a general class-agnostic objectness measure to limit bias toward labeled samples. One issue in open-world detection is that previously unseen objects are often misclassified as known categories or filtered as background by classifiers. To prevent this, we explicitly model the joint distribution of objectness and category labels using variational approximation. However, without sufficient labeled data, minimizing the KL divergence between the estimated posterior and a static normal prior fails to converge. Our theoretical analysis identifies the root cause of this failure and motivates adopting a Gaussian prior with variance dynamically adapted to the estimated posterior as a surrogate. To further reduce misclassification, we introduce an energy-based margin loss that encourages unknown objects to move toward high-density regions of the distribution, thus reducing the uncertainty of unknown detections. Our Open-World OBJectness modeling (OWOBJ) boosts novel object detection, especially in low-data regimes. OWOBJ is a flexible plugin that outperforms baselines in Open-World, Few-Shot, and zero-shot Open-Vocabulary Object Detection.

****

LLaVA-Critic: Learning to Evaluate Multimodal Models

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, Chunyuan Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13618-13628

We introduce LLaVA-Critic, the first open-source large multimodal model (LMM) designed as a generalist evaluator to assess performance across a wide range of multimodal tasks. LLaVA-Critic is trained using a high-quality critic instruction-following dataset that incorporates diverse evaluation criteria and scenarios. Our experiments demonstrate the model's effectiveness in two key areas: (i) LMM-as-a-Judge, where LLaVA-Critic provides reliable evaluation scores, performing on par with or surpassing GPT models on multiple evaluation benchmarks; and (ii) Preference Learning, where it generates reward signals for preference learning, enhancing model alignment capabilities. This work underscores the potential of open-source LMMs in self-critique and evaluation, setting the stage for future research into scalable, superhuman alignment feedback mechanisms for LMMs.

****

VILA-M3: Enhancing Vision-Language Models with Medical Expert Knowledge

Vishwesh Nath, Wenqi Li, Dong Yang, Andriy Myronenko, Mingxin Zheng, Yao Lu, Zhijian Liu, Hongxu Yin, Yee Man Law, Yucheng Tang, Pengfei Guo, Can Zhao, Ziyue Xu, Yufan He, Stephanie Harmon, Benjamin Simon, Greg Heinrich, Stephen Aylward, Marc Edgar, Michael Zephyr, Pavlo Molchanov, Baris Turkbey, Holger Roth, Daguang Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14788-14798

Generalist vision language models (VLMs) have made significant strides in comput
er vision, but they fall short in specialized fields like healthcare, where expe
rt knowledge is essential. Current large multimodal models like Gemini and GPT-4
o are insufficient for medical tasks due to their reliance on memorized internet
 knowledge rather than the nuanced expertise required in healthcare. Meanwhile,
existing medical VLMs (e.g. Med-Gemini) often lack expert consultation as part o
f their design, and many rely on outdated, static datasets that were not created
 with modern, large deep learning models in mind. VLMs are usually trained in th
ree stages: vision pre-training, vision-language pre-training, and instruction f
ine-tuning (IFT). IFT has been typically applied using a mixture of generic and
healthcare data. In contrast, we propose that for medical VLMs, a fourth stage o
f specialized IFT is necessary, which focuses on medical data and includes infor
mation from domain expert models. Domain expert models developed for medical use
 are crucial because they are specifically trained for certain clinical tasks, e
.g. to detect tumors and classify abnormalities through segmentation and classif
ication, which learn fine-grained features of medical data-features that are oft
en too intricate for a VLM to capture effectively. This paper introduces a new f
ramework, VILA-M3, for medical VLMs that utilizes domain knowledge via expert mo
dels. We argue that generic VLM architectures alone are not viable for real-worl
d clinical applications and on-demand usage of domain-specialized expert model k
nowledge is critical for advancing AI in healthcare. Through our experiments, we
 show an improved state-of-the-art (SOTA) performance with an average improvemen
t of ~9% over the prior SOTA model Med-Gemini and ~6% over models trained on the
 specific tasks. Our approach emphasizes the importance of domain expertise in c
reating precise, reliable VLMs for medical applications.
****

Repurposing Pre-trained Video Diffusion Models for Event-based Video Interpolati
on

Jingxi Chen, Brandon Y. Feng, Haoming Cai, Tianfu Wang, Levi Burner, Dehao Yuan,
 Cornelia Fermuller, Christopher A. Metzler, Yiannis Aloimonos; Proceedings of t
he Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12456-12
466

Video Frame Interpolation aims to recover realistic missing frames between obser
ved frames, generating a high-frame-rate video from a low-frame-rate video. Howe
ver, without additional guidance, large motion between frames makes this problem
 ill-posed. Event-based Video Frame Interpolation (EVFI) addresses this challeng
e by using sparse, high-temporal-resolution event measurements as motion guidanc
e. This guidance allows EVFI methods to significantly outperform frame-only meth
ods. However, to date, EVFI methods have relied upon a limited set of paired eve
nt-frame training data, severely limiting their performance and generalization c
apabilities. In this work, we overcome the limited data challenge by adapting p
re-trained video diffusion models trained on internet-scale datasets to EVFI. We
 experimentally validate our approach on real-world EVFI datasets, including a n
ew one we introduce. Our method outperforms existing methods and generalizes acr
oss cameras far better than existing approaches.
****

MotionPRO: Exploring the Role of Pressure in Human MoCap and Beyond

Shenghao Ren, Yi Lu, Jiayi Huang, Jiayi Zhao, He Zhang, Tao Yu, Qiu Shen, Xun Ca
o; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR),
 2025, pp. 27760-27770

Existing human Motion Capture (MoCap) methods mostly focus on the visual similar
ity while neglecting the physical plausibility. As a result, downstream tasks su
ch as driving virtual human in 3D scene or humanoid robots in real world suffer
from issues such as timing drift and jitter, spatial problems like sliding and p
enetration, and poor global trajectory accuracy. In this paper, we revisit human
 MoCap from the perspective of interaction between human body and physical world
 by exploring the role of pressure. Firstly, we construct a large-scale human Mo
tion capture dataset with Pressure, RGB and Optical sensors (named MotionPRO),
which comprises 70 volunteers performing 400 types of motion, encompassing a tot
al of 12.4M pose frames. Secondly, we examine both the necessity and effectivene

ss of the pressure signal through two challenging tasks: (1) pose and trajectory estimation based solely on pressure: We propose a network that incorporates a small kernel decoder and a long-short-term attention module, and proof that pressure could provide accurate global trajectory and plausible lower body pose. (2) pose and trajectory estimation by fusing pressure and RGB: We impose constraints on orthographic similarity along the camera axis and whole-body contact along the vertical axis to enhance the cross-attention strategy to fuse pressure and RGB feature maps. Experiments demonstrate that fusing pressure with RGB features not only significantly improves performance in terms of objective metrics, but also plausibly drives virtual humans (SMPL) in 3D scene. Furthermore, we demonstrate that incorporating physical perception enables humanoid robots to perform more precise and stable actions, which is highly beneficial for the development of embodied artificial intelligence. Project page is available at: https://nju-cite-mocaphumanoid.github.io/MotionPRO/
****

DiffVsgg: Diffusion-Driven Online Video Scene Graph Generation
Mu Chen, Liulei Li, Wenguan Wang, Yi Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29161-29172
Top-leading solutions for Video Scene Graph Generation (VSGG) typically adopt an offline pipeline.Though demonstrating promising performance, they remain unable to handle real-time video streams and consume large GPU memory. Moreover, these approaches fall short in temporal reasoning, merely aggregating frame-level predictions over a temporal context. In response, we introduce DiffVsgg, an online VSGG solution that frames this task as an iterative scene graph update problem. Drawing inspiration from Latent Diffusion Models (LDMs) which generate images via denoising a latent feature embedding, we unify the decoding of object classification, bounding box regression, and graph generation three tasks using one shared feature embedding. Then, given an embedding containing unified features of object pairs, we conduct a step-wise Denoising on it within LDMs, so as to deliver a clean embedding which clearly indicates the relationships between objects.This embedding then serves as the input to task-specific heads for object classification, scene graph generation, etc. DiffVsgg further facilitates continuous temporal reasoning, where predictions for subsequent frames leverage results of past frames as the conditional inputs of LDMs, to guide the reverse diffusion process for current frames.Extensive experiments on three setups of Action Genome demonstrate the superiority of DiffVsgg.
****

Large-scale Multi-view Tensor Clustering with Implicit Linear Kernels
Jiyuan Liu, Xinwang Liu, Chuankun Li, Xinhang Wan, Hao Tan, Yi Zhang, Weixuan Liang, Qian Qu, Yu Feng, Renxiang Guan, Ke Liang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20727-20736
Multi-view clustering is a long-standing hot topic in machine learning communities, due to its capability of integrating data information from multiple sources and modalities. By utilizing tensor Singular Value Decomposition (t-SVD) technique with the tensor rotation trick, recent advances have achieved remarkable improvements on clustering performance. However, we find this is attributed to the inadvertent use of sequential information of sorted data samples, i.e. inadvertent label use, which violates the unsupervised learning setting. On the other hand, existing large-scale approaches are mostly developed on the basis of matrix factorization or anchor techniques, thereby fail to consider the similarities among all data samples, preventing from further performance improvement. To address the above issues, we first analyze the tensor rotation trick and recommend to remove it from tensor clustering. On its basis, a novel large-scale multi-view tensor clustering method is developed by incorporating the pair-wise similarities with implicit linear kernel function. To solve the resultant optimization problem, we design an efficient algorithm of linear complexity.Moreover, extensive experiments are conducted and corresponding results well support the aforementioned finding and validate the effectiveness and efficiency of the proposed method.
****

Generalized Diffusion Detector: Mining Robust Features from Diffusion Models for

Domain-Generalized Detection
Boyong He, Yuxiang Ji, Qianwen Ye, Zhuoyue Tan, Liaoni Wu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9921-9932
Domain generalization (DG) for object detection aims to enhance detectors' performance in unseen scenarios. This task remains challenging due to complex variations in real-world applications. Recently, diffusion models have demonstrated remarkable capabilities in diverse scene generation, which inspires us to explore their potential for improving DG tasks. Instead of generating images, our method extracts multi-step intermediate features during the diffusion process to obtain domain-invariant features for generalized detection. Furthermore, we propose an efficient knowledge transfer framework that enables detectors to inherit the generalization capabilities of diffusion models through feature and object-level alignment, without increasing inference time. We conduct extensive experiments on six challenging DG benchmarks. The results demonstrate that our method achieves substantial improvements of 14.0% mAP over existing DG approaches across different domains and corruption types. Notably, our method even outperforms most domain adaptation methods without accessing any target domain data. Moreover, the diffusion-guided detectors show consistent improvements of 15.9% mAP on average compared to the baseline. Our work aims to present an effective approach for domain-generalized detection and provide potential insights for robust visual recognition in real-world scenarios.
****

Q-Eval-100K: Evaluating Visual Quality and Alignment Level for Text-to-Vision Content
Zicheng Zhang, Tengchuan Kou, Shushi Wang, Chunyi Li, Wei Sun, Wei Wang, Xiaoyu Li, Zongyu Wang, Xuezhi Cao, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10621-10631
Evaluating text-to-vision content hinges on two crucial aspects: **visual quality** and **alignment**. While significant progress has been made in developing objective models to assess these dimensions, the performance of such models heavily relies on the scale and quality of human annotations. According to **Scaling Law**, increasing the number of human-labeled instances follows a predictable pattern that enhances the performance of evaluation models.Therefore, we introduce a comprehensive dataset designed to **E**valuate **V**isual quality and **A**lignment **L**evel for text-to-vision content (**Q-EVAL-100K**), featuring the largest collection of human-labeled Mean Opinion Scores (MOS) for the mentioned two aspects.The **Q-EVAL-100K** dataset encompasses both text-to-image and text-to-video models, with 960K human annotations specifically focused on visual quality and alignment for 100K instances (60K images and 40K videos). Leveraging this dataset with context prompt, we propose **Q-Eval-Score**, a unified model capable of evaluating both visual quality and alignment with special improvements for handling long-text prompt alignment.Experimental results indicate that the proposed **Q-Eval-Score** achieves superior performance on both visual quality and alignment, with strong generalization capabilities across other benchmarks. These findings highlight the significant value of the **Q-EVAL-100K** dataset. **The data and code will be released** to help promote the generation models.
****

Dual Focus-Attention Transformer for Robust Point Cloud Registration
Kexue Fu, Mingzhi Yuan, Changwei Wang, Weiguang Pang, Jing Chi, Manning Wang, Longxiang Gao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11769-11778
Recently, coarse-to-fine methods for point cloud registration have achieved great success, but few works deeply explore the impact of feature interaction at both coarse and fine scales. By visualizing attention scores and correspondences, we find that existing methods fail to achieve effective feature aggregation at the two scales during the feature interaction. To tackle this issue, we propose a Dual Focus-Attention Transformer framework, which only focuses on points relevant to the current point for feature interaction, avoiding interactions with irrelevant points. For the coarse scale, we design a superpoint focus-attention trans

former guided by sparse keypoints, which are selected from the neighborhood of s uperpoints. For the fine scale, we only perform feature interaction between the point sets that belong to the same superpoint. Experiments show that our method achieve the state-of-the-art performance on three standard benchmarks. The code and pre-trained models will be available at Github.
****

Forming Auxiliary High-confident Instance-level Loss to Promote Learning from La bel Proportions
Tianhao Ma, Han Chen, Juncheng Hu, Yungang Zhu, Ximing Li; Proceedings of the Co mputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20592-20601
Learning from label proportions (LLP), i.e. a challenging weakly-supervised lear ning task, aims to train a classifier by using bags of instances and the proport ions of classes within bags, rather than annotated labels for each instance. Bey ond the traditional bag-level loss, the mainstream methodology of LLP is to inco rporate an auxiliary instance-level loss with pseudo-labels formed by prediction s. Unfortunately, we empirically observed that the pseudo-labels are often inacc urate and even meaningless, especially for the scenarios with large bag sizes, h urting the classifier induction. To alleviate this problem, we suggest a novel L LP method, namely Learning Label Proportions with Auxiliary High-confident Insta nce-level Loss (L^2P-AHIL). Specifically, we propose a dual entropy-based weight (DEW) method to adaptively measure the confidences of pseudo-labels. It simulta neously emphasizes accurate predictions at the bag level and avoids smoothing pr edictions, which tend to be meaningless. We then form high-confident instance-le vel loss with DEW, and jointly optimize it with the bag-level loss in a self-tra ining manner. The experimental results on benchmark datasets show that L^2P-AHIL can surpass the existing baseline methods, and the performance gain can be more significant as the bag size increases. The implementation of our method is avai lable at https://github.com/TianhaoMa5/LLP-AHIL.
****

Progress-Aware Video Frame Captioning
Zihui Xue, Joungbin An, Xitong Yang, Kristen Grauman; Proceedings of the Compute r Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13639-13650
While image captioning provides isolated descriptions for individual images, and video captioning offers one single narrative for an entire video clip, our work explores an important middle ground: progress-aware video captioning at the fra me level. This novel task aims to generate temporally fine-grained captions that not only accurately describe each frame but also capture the subtle progression of actions throughout a video sequence. Despite the strong capabilities of exis ting leading vision language models, they often struggle to discern the nuances of frame-wise differences. To address this, we propose ProgressCaptioner, a capt ioning model designed to capture the fine-grained temporal dynamics within an ac tion sequence. Alongside, we develop the FrameCap dataset to support training an d the FrameCapEval benchmark to assess caption quality. The results demonstrate that ProgressCaptioner significantly surpasses leading captioning models, produc ing precise captions that accurately capture action progression and set a new st andard for temporal precision in video captioning. Finally, we showcase practica l applications of our approach, specifically in aiding keyframe selection and ad vancing video understanding, highlighting its broad utility.
****

SMTPD: A New Benchmark for Temporal Prediction of Social Media Popularity
Yijie Xu, Bolun Zheng, Wei Zhu, Hangjia Pan, Yuchen Yao, Ning Xu, Anan Liu, Quan Zhang, Chenggang Yan; Proceedings of the Computer Vision and Pattern Recognitio n Conference (CVPR), 2025, pp. 18847-18857
Social media popularity prediction task aims to predict the popularity of posts on social media platforms, which has a positive driving effect on application sc enarios such as content optimization, digital marketing and online advertising. Though many studies have made significant progress, few of them pay much attenti on to the integration between popularity prediction with temporal alignment. In this paper, with exploring YouTube's multilingual and multi-modal content, we co nstruct a new social media temporal popularity prediction benchmark, namely SMTP

D, and suggest a baseline framework for temporal popularity prediction. Through data analysis and experiments, we verify that temporal alignment and early popularity play crucial roles in social media popularity prediction for not only deepening the understanding of temporal dynamics of popularity in social media but also offering a suggestion about developing more effective prediction models in this field. Code is available at https://github.com/zhuwei321/SMTPD

****

Enhancing Dance-to-Music Generation via Negative Conditioning Latent Diffusion Model

Changchang Sun, Gaowen Liu, Charles Fleming, Yan Yan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8321-8330

Conditional diffusion models have gained increasing attention since their impressive results for cross-modal synthesis, where the strong alignment between conditioning input and generated output can be achieved by training a time-conditioned U-Net augmented with cross-attention mechanism. In this paper, we focus on the problem of generating music synchronized with rhythmic visual cues of the given dance video. Considering that bi-directional guidance is more beneficial for training a diffusion model, we propose to enhance the quality of generated music and its synchronization with dance videos by adopting both positive rhythmic information and negative ones (PN-Diffusion) as conditions, where a dual diffusion and reverse processes is devised. Specifically, to train a sequential multi-modal U-Net structure, PN-Diffusion consists of a noise prediction objective for positive conditioning and an additional noise prediction objective for negative conditioning. To accurately define and select both positive and negative conditioning, we ingeniously utilize temporal correlations in dance videos, capturing positive and negative rhythmic cues by playing them forward and backward, respectively. Through subjective and objective evaluations of input-output correspondence in terms of dance-music beat alignment and the quality of generated music, experimental results on the AIST++ and TikTok dance video datasets demonstrate that our model outperforms SOTA dance-to-music generation models.

****

Neuro-Symbolic Evaluation of Text-to-Video Models using Formal Verification

S P Sharan, Minkyu Choi, Sahil Shah, Harsh Goel, Mohammad Omama, Sandeep Chinchali; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8395-8405

Recent advancements in text-to-video models such as Sora, Gen-3, MovieGen, and CogVideoX are pushing the boundaries of synthetic video generation, with adoption seen in fields like robotics, autonomous driving, and entertainment. As these models become prevalent, various metrics and benchmarks have emerged to evaluate the quality of the generated videos. However, these metrics emphasize visual quality and smoothness, neglecting temporal fidelity and text-to-video alignment, which are crucial for safety-critical applications. To address this gap, we introduce NeuS-V, a novel synthetic video evaluation metric that rigorously assesses text-to-video alignment using neuro-symbolic formal verification techniques. Our approach first converts the prompt into a formally defined Temporal Logic (TL) specification and translates the generated video into an automaton representation. Then, it evaluates the text-to-video alignment by formally checking the video automaton against the TL specification. Furthermore, we present a dataset of temporally extended prompts to evaluate state-of-the-art video generation models against our benchmark. We find that NeuS-V demonstrates a higher correlation by over 5x with human evaluations when compared to existing metrics. Our evaluation further reveals that current video generation models perform poorly on these temporally complex prompts, highlighting the need for future work in improving text-to-video generation capabilities. We open-source our benchmark, code, and dataset at https://utaustin-swarmlab.github.io/neusv.

****

Spherical Manifold Guided Diffusion Model for Panoramic Image Generation

Xiancheng Sun, Mai Xu, Shengxi Li, Senmao Ma, Xin Deng, Lai Jiang, Gang Shen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5824-5834

Panoramic image essentially acts as a pivotal role in emerging virtual reality a
nd augmented reality scenarios; however, the generation of panoramic images are
essentially challenging due to the intrinsic spherical geometry and spherical di
stortions caused by equirectangular projection (ERP). To address this, we start
from the very basics of S^2 manifold inherent to panoramic images, and propose a
 novel spherical manifold convolution (SMConv) on S^2 manifold. Based on the SMC
onv operation, we propose a spherical manifold guided diffusion (SMGD) model for
 text-conditioned panoramic image generation, which can well accommodate the sph
erical geometry during generation. We further develop a novel evaluation method
by calculating grouped Frechet inception distance (FID) on cube-map projections,
 which can well reflect the quality of generated panoramic images, compared to e
xisting methods that randomly crop ERP-distorted content. Experiment results dem
onstrate that our SMGD model achieves the state-of-the-art generation quality an
d accuracy, whilst retaining the shortest sampling time in the text-conditioned
panoramic image generation task. Codes are publicly available at https://github.
com/chronos123/SMGD.
****

Learning on Model Weights using Tree Experts
Eliahu Horwitz, Bar Cavia, Jonathan Kahana, Yedid Hoshen; Proceedings of the Com
puter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20468-20478
The number of publicly available models is rapidly increasing, yet most remain u
ndocumented. Users looking for suitable models for their tasks must first determ
ine what each model does. Training machine learning models to infer missing docu
mentation directly from model weights is challenging, as these weights often con
tain significant variation unrelated to model functionality (denoted nuisance).
Here, we identify a key property of real-world models: most public models belong
 to a small set of Model Trees, where all models within a tree are fine-tuned fr
om a common ancestor (e.g., a foundation model). Importantly, we find that withi
n each tree there is less nuisance variation between models. Concretely, while l
earning across Model Trees requires complex architectures, even a linear classif
ier trained on a single model layer often works within trees. While effective, t
hese linear classifiers are computationally expensive, especially when dealing w
ith larger models that have many parameters. To address this, we introduce Probi
ng Experts (ProbeX), a theoretically motivated and lightweight method. Notably,
ProbeX is the first probing method specifically designed to learn from the weigh
ts of a single hidden model layer. We demonstrate the effectiveness of ProbeX by
 predicting the categories in a model's training dataset based only on its weigh
ts. Excitingly, ProbeX can map the weights of Stable Diffusion into a weight-lan
guage embedding space, enabling model search via text, i.e., zero-shot model cla
ssification.
****

Rethinking Query-based Transformer for Continual Image Segmentation
Yuchen Zhu, Cheng Shi, Dingyou Wang, Jiajin Tang, Zhengxuan Wei, Yu Wu, Guanbin
Li, Sibei Yang; Proceedings of the Computer Vision and Pattern Recognition Confe
rence (CVPR), 2025, pp. 4595-4606
Class-incremental/Continual image segmentation (CIS) aims to train an image segm
enter in stages, where the set of available categories differs at each stage. To
 leverage the built-in objectness of query-based transformers, which mitigates c
atastrophic forgetting of mask proposals, current methods often decouple mask ge
neration from the continual learning process. This study, however, identifies tw
o key issues with decoupled frameworks: loss of plasticity and heavy reliance on
 input data order. To address these, we conduct an in-depth investigation of the
 built-in objectness and find that highly aggregated image features provide a sh
ortcut for queries to generate masks through simple feature alignment. Based on
this, we propose SimCIS, a simple yet powerful baseline for CIS. Its core idea i
s to directly select image features for query assignment, ensuring "perfect alig
nment" to preserve objectness, while simultaneously allowing queries to select n
ew classes to promote plasticity. To further combat catastrophic forgetting of c
ategories, we introduce cross-stage consistency in selection and an innovative "
visual query"-based replay mechanism. Experiments demonstrate that SimCIS consis

tently outperforms state-of-the-art methods across various segmentation tasks, s
ettings, splits, and input data orders. All models and codes will be made public
ly available at https://github.com/SooLab/SimCIS.
****

Image Reconstruction from Readout-Multiplexed Single-Photon Detector Arrays
Shashwath Bharadwaj, Ruangrawee Kitichotkul, Akshay Agarwal, Vivek K Goyal; Proc
eedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025,
pp. 11406-11415
Readout multiplexing is a promising solution to overcome hardware limitations an
d data bottlenecks in imaging with single-photon detectors. Conventional multipl
exed readout processing creates an upper bound on photon counts at a very fine t
ime scale, where frames with multiple detected photons must either be discarded
or allowed to introduce significant bias. We formulate multiphoton coincidence r
esolution as an inverse imaging problem and introduce a solution framework to pr
obabilistically resolve the spatial locations of photon incidences. Specifically
, we develop a theoretical abstraction of row--column multiplexing and a model o
f photon events that make readouts ambiguous. Using this, we propose a novel est
imator that spatially resolves up to four coincident photons. Monte Carlo simula
tions show that our proposed method increases the peak signal-to-noise ratio (PS
NR) of reconstruction by 3 to 4 dB compared to conventional methods under optima
l incident flux conditions. Additionally, this method reduces the required numbe
r of readout frames to achieve the same mean-squared error as other methods by a
 factor of 4. Finally, our method matches the Cramer-Rao bound for detection pr
obability estimation for a wider range of incident flux values compared to conve
ntional methods. While demonstrated for a specific detector type and readout arc
hitecture, this method can be extended to more general multiplexing with differe
nt detector models.
****

Towards Smart Point-and-Shoot Photography
Jiawan Li, Fei Zhou, Zhipeng Zhong, Jiongzhi Lin, Guoping Qiu; Proceedings of th
e Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28242-282
51
Hundreds of millions of people routinely take photos using their smartphones as
point and shoot (PAS) cameras, yet very few would have the photography skills to
 compose a good shot of a scene. While traditional PAS cameras have built-infunc
tions to ensure a photo is well focused and has the right brightness, they canno
t tell the users how to compose the best shot of a scene. In this paper, we pres
ent a first of its kind smart point and shoot (SPAS) system to help users to tak
e good photos. Our SPAS proposes to help users to compose a good shot of a scene
 by automatically guiding the users to adjust the camera pose live on the scene.
 We first constructed a large dataset containing 320K images with camera pose in
formation from 4000 scenes. We then developed an innovative CLIP-based Compositi
on Quality Assessment (CCQA) model to assign pseudo labels to these images. The
CCQA introduces a unique learnable text embedding technique to learn continuous
word embeddings capable of discerning subtle visual quality differences in the r
ange covered by five levels of quality description words bad, poor, fair, good,
 perfect . And finally we have developed a camera pose adjustment model (CPAM) w
hich first determines if the current view can be further improved and if so it o
utputs the adjust suggestion in the form of two camera pose adjustment angles. T
he two tasks of CPAM make decisions in a sequential manner and each involves dif
ferent sets of training samples, we have developed a mixture-of-experts model wi
th a gated loss function to train the CPAM in an end-to-end manner. We will pres
ent extensive results to demonstrate the performances of our SPAS system using p
ublicly available image composition datasets.
****

SlideChat: A Large Vision-Language Assistant for Whole-Slide Pathology Image Und
erstanding
Ying Chen, Guoan Wang, Yuanfeng Ji, Yanjun Li, Jin Ye, Tianbin Li, Ming Hu, Rong
shan Yu, Yu Qiao, Junjun He; Proceedings of the Computer Vision and Pattern Reco
gnition Conference (CVPR), 2025, pp. 5134-5143

Despite the progress made by multimodal large language models (MLLMs) in computational pathology, they remain limited by a predominant focus on patch-level analysis, missing essential contextual information at the whole-slide level. The lack of large-scale instruction datasets and the gigapixel scale of whole slide images (WSIs) pose significant developmental challenges. In this paper, we present SlideChat, the first vision-language assistant capable of understanding gigapixel whole-slide images, exhibiting excellent multimodal conversational capability and response complex instruction across diverse pathology scenarios. To support its development, we created SlideInstruction, the largest instruction-following dataset for WSIs consisting of 4.2K WSI captions and 176K VQA pairs with multiple categories. Furthermore, we propose SlideBench, a multimodal benchmark that incorporates captioning and VQA tasks to assess SlideChat's capabilities in various settings such as microscopy, diagnosis and clinical. Compared to both general and specialized MLLMs, SlideChat exhibits exceptional capabilities, achieving state-of-the-art performance on 18 of 22 tasks. For example, it achieved an overall accuracy of 81.17% on SlideBench-VQA (TCGA), and 54.15% on SlideBench-VQA (BCNB). Our code, data, and model is publicly accessible at https://uni-medical.github.io/SlideChat.github.io.

****

Prototype-Based Image Prompting for Weakly Supervised Histopathological Image Segmentation

Qingchen Tang, Lei Fan, Maurice Pagnucco, Yang Song; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30271-30280

Weakly supervised image segmentation with image-level labels has drawn attention due to the high cost of pixel-level annotations. Traditional methods using Class Activation Maps (CAMs) often highlight only the most discriminative regions, leading to incomplete masks. Recent approaches that introduce textual information struggle with histopathological images due to inter-class homogeneity and intra-class heterogeneity. In this paper, we propose a prototype-based image prompting framework for histopathological image segmentation. It constructs an image bank from the training set using clustering, extracting multiple prototype features per class to capture intra-class heterogeneity. By designing a matching loss between input features and class-specific prototypes using contrastive learning, our method addresses inter-class homogeneity and guides the model to generate more accurate CAMs. Experiments on four datasets (LUAD-HistoSeg, BCSS-WSSS, GCSS, and BCSS) show that our method outperforms existing weakly supervised segmentation approaches, setting new benchmarks in histopathological image segmentation.

****

Towards Transformer-Based Aligned Generation with Self-Coherence Guidance

Shulei Wang, Wang Lin, Hai Huang, Hanting Wang, Sihang Cai, WenKang Han, Tao Jin, Jingyuan Chen, Jiacheng Sun, Jieming Zhu, Zhou Zhao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18455-18464

We introduce a novel, training-free approach for enhancing alignment in Transformer-based Text-Guided Diffusion Models (TGDMs). Existing TGDMs often struggle to generate semantically aligned images, particularly when dealing with complex text prompts or multi-concept attribute binding challenges. Previous U-Net-based methods primarily optimized the latent space, but their direct application to Transformer-based architectures has shown limited effectiveness. Our method addresses these challenges by directly optimizing cross-attention maps during the generation process. Specifically, we introduce Self-Coherence Guidance, a method that dynamically refines attention maps using masks derived from previous denoising steps, ensuring precise alignment without additional training. To validate our approach, we constructed more challenging benchmarks for evaluating coarse-grained attribute binding, fine-grained attribute binding, and style binding. Experimental results demonstrate the superior performance of our method, significantly surpassing other state-of-the-art methods across all evaluated tasks. Our code is available at https://scg-diffusion.github.io/scg-diffusion.

****

Accurate Scene Text Recognition with Efficient Model Scaling and Cloze Self-Distillation

Andrea Maracani, Savas Ozkan, Sijun Cho, Hyowon Kim, Eunchung Noh, Jeongwon Min, Cho Jung Min, Dookun Park, Mete Ozay; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14516-14526
Scaling architectures have been proven effective for improving Scene Text Recognition (STR), but the individual contribution of vision encoder and text decoder scaling remain under-explored. In this work, we present an in-depth empirical analysis and demonstrate that, contrary to previous observations, scaling the decoder yields significant performance gains, always exceeding those achieved by encoder scaling alone. We also identify label noise as a key challenge in STR, particularly in real-world data, which can limit the effectiveness of STR models. To address this, we propose Cloze Self-Distillation (CSD), a method that mitigates label noise by distilling a student model from context-aware soft predictions and pseudolabels generated by a teacher model. Additionally, we enhance the decoder architecture by introducing differential cross-attention for STR. Our methodology achieves state-of-the-art performance on 10 out of 11 benchmarks using only real data, while significantly reducing the parameter size and computational costs.
****

DART: Disease-aware Image-Text Alignment and Self-correcting Re-alignment for Trustworthy Radiology Report Generation
Sang-Jun Park, Keun-Soo Heo, Dong-Hee Shin, Young-Han Son, Ji-Hye Oh, Tae-Eui Kam; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15580-15589
The automatic generation of radiology reports has emerged as a promising solution to reduce a time-consuming task and accurately capture critical disease-relevant findings in X-ray images. Previous approaches for radiology report generation have shown impressive performance. However, there remains significant potential to improve accuracy by ensuring that retrieved reports contain disease-relevant findings similar to those in the X-ray images and by refining generated reports. In this study, we propose a Disease-aware image-text Alignment and self-correcting Re-alignment for Trustworthy radiology report generation (DART) framework. In the first stage, we generate initial reports based on image-to-text retrieval with disease-matching, embedding both images and texts in a shared embedding space through contrastive learning. This approach ensures the retrieval of reports with similar disease-relevant findings that closely align with the input X-ray images. In the second stage, we further enhance the initial reports by introducing a self-correction module that re-aligns them with the X-ray images. Our proposed framework achieves state-of-the-art results on two widely used benchmarks, surpassing previous approaches in both report generation and clinical efficacy metrics, thereby enhancing the trustworthiness of radiology reports.
****

On the Consistency of Video Large Language Models in Temporal Comprehension
Minjoon Jung, Junbin Xiao, Byoung-Tak Zhang, Angela Yao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13713-13722
Video large language models (Video-LLMs) can temporally ground language queries and retrieve video moments. Yet, such temporal comprehension capabilities are neither well-studied nor understood. So we conduct a study on prediction consistency -- a key indicator for robustness and trustworthiness of temporal grounding. After the model identifies an initial moment within the video content, we apply a series of probes to check if the model's responses align with this initial grounding as an indicator of reliable comprehension. Our results reveal that current Video-LLMs are sensitive to variations in video contents, language queries, and task settings, unveiling severe deficiencies in maintaining consistency. We further explore common prompting and instruction-tuning methods as potential solutions, but find that their improvements are often unstable. To that end, we propose event temporal verification tuning that explicitly accounts for consistency, and demonstrate significant improvements for both grounding and consistency. Our data and code are open-sourced at https://github.com/minjoong507/Consistency-of-Video-LLM.
****

Mitigating the Human-Robot Domain Discrepancy in Visual Pre-training for Robotic Manipulation
Jiaming Zhou, Teli Ma, Kun-Yu Lin, Zifan Wang, Ronghe Qiu, Junwei Liang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22551-22561
Learning generalizable visual representations across different embodied environments is essential for effective robotic manipulation in real-world scenarios. However, the limited scale and diversity of robot demonstration data pose a significant challenge. Recent research has explored leveraging large-scale human activity data for pre-training, but the substantial morphological differences between humans and robots introduce a significant human-robot domain discrepancy, hindering the generalization of these models to downstream manipulation tasks. To overcome this, we propose a novel adaptation paradigm that leverages readily available paired human-robot video data to bridge the domain gap. Our method employs a human-robot contrastive alignment loss to align the semantics of human and robot videos, adapting pre-trained models to the robot domain in a parameter-efficient manner. Experiments on 20 simulated tasks across two different benchmarks and five real-world tasks demonstrate significant improvements. These results span both single-task and language-conditioned multi-task settings, evaluated using two different pre-trained models. Compared to existing pre-trained models, our adaptation method improves the average success rate by over 7% across multiple tasks on both simulated benchmarks and real-world evaluations.
****
Less is More: Efficient Model Merging with Binary Task Switch
Biqing Qi, Fangyuan Li, Zhen Wang, Junqi Gao, Dong Li, Peng Ye, Bowen Zhou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15265-15274
As an effective approach to equip models with multi-task capabilities without additional training, model merging has garnered significant attention. However, existing merging methods face challenges of redundant parameter conflicts and the excessive storage burden of fine-tuned parameters. In this work, through controlled experiments, we reveal that for fine-tuned task vectors, only those parameters with magnitudes above a certain threshold contribute positively to the task, exhibiting a pulse-like characteristic. We then attempt leveraging this pulse-like characteristic to binarize the task vectors and reduce storage overhead. Further controlled experiments show that the binarized task vectors incur almost no decrease in fine-tuning and merging performance, and even exhibit stronger performance improvements as the proportion of redundant parameters increases. Based on these insights, we propose Task Switch (T-Switch), which decomposes task vectors into three components: 1) an activation switch instantiated by a binarized mask vector, 2) a polarity switch instantiated by a binarized sign vector, and 3) a scaling knob instantiated by a scalar coefficient. By storing task vectors in a binarized form, T-Switch alleviates parameter conflicts while ensuring efficient task parameter storage. Furthermore, to enable automated switch combination in T-Switch, we further introduce Auto-Switch, which enables training-free switch combination via retrieval from a small query set. Experiments indicate that our methods achieve significant performance improvements over existing baselines, requiring only 1-3% of the storage space of full-precision parameters.
****
One-Minute Video Generation with Test-Time Training
Karan Dalal, Daniel Koceja, Jiarui Xu, Yue Zhao, Shihao Han, Ka Chun Cheung, Jan Kautz, Yejin Choi, Yu Sun, Xiaolong Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17702-17711
Transformers today still struggle to generate one-minute videos because self-attention layers are inefficient for long context. Alternatives such as Mamba layers struggle to produce coherent scenes because their hidden states are small and less expressive. We experiment with Test-Time Training (TTT) layers, whose hidden states themselves can be neural networks, therefore larger and more expressive. Adding TTT layers into a pre-trained Transformer enables it to generate one-minute videos from text storyboards. We curate a dataset based on Tom and Jerry ca

rtoons as a proof-of-concept benchmark. Compared to baselines such as Mamba 2, G ated DeltaNet, and sliding-window attention layers, TTT layers generate much mor e coherent videos that tell complete stories, leading by 34 Elo points in a huma n evaluation of 100 videos per method. Although promising, our results are still limited in physical realism, and the efficiency of our implementation can be fu rther improved. Sample videos, code and annotations are available at: https://te st-time-training.github.io/video-dit
****

InteractionMap: Improving Online Vectorized HDMap Construction with Interaction
Kuang Wu, Chuan Yang, Zhanbin Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17176-17186
Vectorized high-definition (HD) maps are essential for an autonomous driving sys tem. Recently, state-of-the-art map vectorization methods are mainly based on DE TR-like framework to generate HD maps in an end-to-end manner. In this paper, we propose InteractionMap, which improves previous map vectorization methods by fu lly leveraging local-to-global information interaction in both time and space. F irstly, we explore enhancing DETR-like detectors by explicit position relation p rior from point-level to instance-level, since map elements contain strong shape priors. Secondly, we propose a key-frame-based hierarchical temporal fusion mod ule, which interacts temporal information from local to global. Lastly, the sepa rate classification branch and regression branch lead to the problem of misalign ment in the output distribution. We interact semantic information with geometric information by introducing a novel geometric-aware classification loss in optim ization and a geometric-aware matching cost in label assignment. InteractionMap achieves state-of-the-art performance on both nuScenes and Argoverse2 benchmarks .
****

Text-guided Sparse Voxel Pruning for Efficient 3D Visual Grounding
Wenxuan Guo, Xiuwei Xu, Ziwei Wang, Jianjiang Feng, Jie Zhou, Jiwen Lu; Proceedi ngs of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3666-3675
In this paper, we propose an efficient multi-level convolution architecture for 3D visual grounding. Conventional methods are difficult to meet the requirements of real-time inference due to the two-stage or point-based architecture. Inspir ed by the success of multi-level fully sparse convolutional architecture in 3D o bject detection, we aim to build a new 3D visual grounding framework following t his technical route. However, as in 3D visual grounding task the 3D scene repres entation should be deeply interacted with text features, sparse convolution-base d architecture is inefficient for this interaction due to the large amount of vo xel features. To this end, we propose text-guided pruning (TGP) and completion-b ased addition (CBA) to deeply fuse 3D scene representation and text features in an efficient way by gradual region pruning and target completion. Specifically, TGP iteratively sparsifies the 3D scene representation and thus efficiently inte racts the voxel features with text features by cross-attention. To mitigate the affect of pruning on delicate geometric information, CBA adaptively fixes the ov er-pruned region by voxel completion with negligible computational overhead. Com pared with previous single-stage methods, our method achieves top inference spee d and surpasses previous fastest method by 100% FPS. Our method also achieves st ate-of-the-art accuracy even compared with two-stage methods, with +1.13 lead of Acc@0.5 on ScanRefer, and +2.6 and +3.2 leads on NR3D and SR3D respectively. Th e code is available at https://github.com/GWxuan/TSP3D.
****

ROCKET-1: Mastering Open-World Interaction with Visual-Temporal Context Promptin g
Shaofei Cai, Zihao Wang, Kewei Lian, Zhancun Mu, Xiaojian Ma, Anji Liu, Yitao Li ang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12122-12131
Vision-language models (VLMs) have excelled in multimodal tasks, but adapting th em to embodied decision-making in open-world environments presents challenges. O ne critical issue is bridging the gap between discrete entities in low-level obs

ervations and the abstract concepts required for effective planning. A common so lution is building hierarchical agents, where VLMs serve as high-level reasoners that break down tasks into executable sub-tasks, typically specified using lang uage. However, language suffers from the inability to communicate detailed spati al information. We propose visual-temporal context prompting, a novel communicat ion protocol between VLMs and policy models. This protocol leverages object segm entation from past observations to guide policy-environment interactions. Using this approach, we train ROCKET-2, a low-level policy that predicts actions based on concatenated visual observations and segmentation masks, supported by real-t ime object tracking from SAM2. Our method unlocks the potential of VLMs, enablin g them to tackle complex tasks that demand spatial reasoning. Experiments in Mi necraft show that our approach enables agents to achieve previously unattainable tasks, with a 76% absolute improvement in open-world interaction performance. C odes are available at https://craftjarvis.github.io/ROCKET-1.

****

Common3D: Self-Supervised Learning of 3D Morphable Models for Common Objects in Neural Feature Space

Leonhard Sommer, Olaf Dünkel, Christian Theobalt, Adam Kortylewski; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6468 -6479

3D morphable models (3DMMs) are a powerful tool to represent the possible shapes and appearances of an object category. Given a single test image, 3DMMs can be used to solve various tasks, such as predicting the 3D shape, pose, semantic cor respondence, and instance segmentation of an object. Unfortunately, 3DMMs are on ly available for very few object categories that are of particular interest, lik e faces or human bodies, as they require a demanding 3D data acquisition and cat egory-specific training process. In contrast, we introduce a new method, Common3 D, that learns 3DMMs of common objects in a fully self-supervised manner from a collection of object-centric videos. For this purpose, our model represents obje cts as a learned 3D template mesh and a deformation field that is parameterized as an image-conditioned neural network. Different from prior works, Common3D rep resents the object appearance with neural features instead of RGB colors, which enables the learning of more generalizable representations through an abstractio n from pixel intensities. Importantly, we train the appearance features using a contrastive objective by exploiting the correspondences defined through the defo rmable template mesh. This leads to higher quality correspondence features compa red to related works and a significantly improved model performance at estimatin g 3D object pose and semantic correspondence. Common3D is the first completely s elf-supervised method that can solve various vision tasks in a zero-shot manner. github.com/GenIntel/common3d.

****

RLAIF-V: Open-Source AI Feedback Leads to Super GPT-4V Trustworthiness

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganq u Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, Maosong Sun; Proceedings of the Computer Vision and Pattern Recog nition Conference (CVPR), 2025, pp. 19985-19995

Traditional feedback learning for hallucination reduction relies on labor-intens ive manual labeling or expensive proprietary models.This leaves the community wi thout foundational knowledge about how to build high-quality feedback with open -source MLLMs.In this work, we introduce RLAIF-V, a novel framework that aligns MLLMs in a fully open-source paradigm. RLAIF-V maximally explores open-source ML LMs from two perspectives, including high-quality feedback data generation for p reference learning and self-feedback guidance for inference-time scaling.Extensi ve experiments on seven benchmarks in both automatic and human evaluation show t hat RLAIF-V substantially enhances the trustworthiness of models at both prefere nce learning and inference time. RLAIF-V 7B reduces object hallucination by 80.7 % and overall hallucination by 33.7%. Remarkably, RLAIF-V 12B further reveals th e self-alignment potential of open-source MLLMs, where the model can learn from feedback of itself to achieve super GPT-4V trustworthiness.

****

ECVC: Exploiting Non-Local Correlations in Multiple Frames for Contextual Video Compression
Wei Jiang, Junru Li, Kai Zhang, Li Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7331-7341
In Learned Video Compression (LVC), improving inter prediction, such as enhancing temporal context mining and mitigating accumulated errors, is crucial for boosting rate-distortion performance. Existing LVCs mainly focus on mining the temporal movements while neglecting non-local correlations among frames. Additionally, current contextual video compression models use a single reference frame, which is insufficient for handling complex movements. To address these issues, we propose leveraging non-local correlations across multiple frames to enhance temporal priors, significantly boosting rate-distortion performance. To mitigate error accumulation, we introduce a partial cascaded fine-tuning strategy that supports fine-tuning on full-length sequences with constrained computational resources. This method reduces the train-test mismatch in sequence lengths and significantly decreases accumulated errors. Based on the proposed techniques, we present a video compression scheme ECVC. Experiments demonstrate that our ECVC achieves state-of-the-art performance, reducing 10.5% and 11.5% more bit-rates than previous SOTA method DCVC-FM over VTM-13.2 low delay B (LDB) under the intra period (IP) of 32 and -1, respectively. Our code is available at https://github.com/JiangWeibeta/ECVC.
****
LinGen: Towards High-Resolution Minute-Length Text-to-Video Generation with Linear Computational Complexity
Hongjie Wang, Chih-Yao Ma, Yen-Cheng Liu, Ji Hou, Tao Xu, Jialiang Wang, Felix Juefei-Xu, Yaqiao Luo, Peizhao Zhang, Tingbo Hou, Peter Vajda, Niraj K. Jha, Xiaoliang Dai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2578-2588
Text-to-video generation enhances content creation but is highly computationally intensive: The computational cost of Diffusion Transformers (DiTs) scales quadratically in the number of pixels. This makes minute-length video generation extremely expensive, limiting most existing models to generating videos of only 10-20 seconds length. We propose a Linear-complexity text-to-video Generation (LinGen) framework whose cost scales linearly in the number of pixels. For the first time, LinGen enables high-resolution minute-length video generation on a single GPU without compromising quality. It replaces the computationally-dominant and quadratic-complexity block, self-attention, with a linear-complexity block called MATE, which consists of an MA-branch and a TE-branch. The MA-branch targets short-to-long-range correlations, combining a bidirectional Mamba2 block with our token rearrangement method, Rotary Major Scan, and our review tokens developed for long video generation. The TE-branch is a novel TEmporal Swin Attention block that focuses on temporal correlations between adjacent tokens and medium-range tokens. The MATE block addresses the adjacency preservation issue of Mamba and improves the consistency of generated videos significantly. Experimental results show that LinGen outperforms DiT (with a 75.6% win rate) in video quality with up to 15x(11.5x) FLOPs (latency) reduction. Furthermore, both automatic metrics and human evaluation demonstrate our LinGen-4B yields comparable video quality to state-of-the-art models (with a 50.5%, 52.1%, 49.1% win rate with respect to Gen-3, LumaLabs, and Kling, respectively). This paves the way to hour-length movie generation and real-time interactive video generation. Project website: https://lineargen.github.io/.
****
EditSplat: Multi-View Fusion and Attention-Guided Optimization for View-Consistent 3D Scene Editing with 3D Gaussian Splatting
Dong In Lee, Hyeongcheol Park, Jiyoung Seo, Eunbyung Park, Hyunje Park, Ha Dam Baek, Sangheon Shin, Sangmin Kim, Sangpil Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11135-11145
Recent advancements in 3D editing have highlighted the potential of text-driven methods in real-time, user-friendly AR/VR applications. However, current methods rely on 2D diffusion models without adequately considering multi-view informati

on, resulting in multi-view inconsistency. While 3D Gaussian Splatting (3DGS) si gnificantly improves rendering quality and speed, its 3D editing process encount ers difficulties with inefficient optimization, as pre-trained Gaussians retain excessive source information, hindering optimization. To address these limitatio ns, we propose EditSplat, a novel 3D editing framework that integrates Multi-vie w Fusion Guidance (MFG) and Attention-Guided Trimming (AGT). Our MFG ensures mul ti-view consistency by incorporating essential multi-view information into the d iffusion process, leveraging classifier-free guidance from the text-to-image dif fusion model and the geometric properties of 3DGS. Additionally, our AGT leverag es the explicit representation of 3DGS to selectively prune and optimize 3D Gaus sians, enhancing optimization efficiency and enabling precise, semantically rich local edits. Through extensive qualitative and quantitative evaluations, EditSp lat achieves superior multi-view consistency and editing quality over existing m ethods, significantly enhancing overall efficiency.

****

SpatialCLIP: Learning 3D-aware Image Representations from Spatially Discriminati ve Language

Zehan Wang, Sashuai Zhou, Shaoxuan He, Haifeng Huang, Lihe Yang, Ziang Zhang, Xi ze Cheng, Shengpeng Ji, Tao Jin, Hengshuang Zhao, Zhou Zhao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29656-29666

Contrastive Language-Image Pre-training (CLIP) learns robust visual models throu gh language supervision, making it a crucial visual encoding technique for vario us applications. However, CLIP struggles with comprehending spatial concepts in images, potentially restricting the spatial intelligence of CLIP-based AI system s. In this work, we propose SpatialCLIP, an enhanced version of CLIP with better spatial understanding capabilities. To capture the intricate 3D spatial relatio nships in images, we improve both "visual model" and "language supervision" of C LIP. Specifically, we design 3D-inspired ViT to replace the standard ViT in CLIP . By lifting 2D image tokens into 3D space and incorporating design insights fro m point cloud networks, our visual model gains greater potential for spatial per ception. Meanwhile, captions with accurate and detailed spatial information are very rare. To explore better language supervision for spatial understanding, we re-caption images and perturb their spatial phrases as negative descriptions, wh ich compels the visual model to seek spatial cues to distinguish these hard nega tive captions. With the enhanced visual model, we introduce SpatialLLaVA, follow ing the same LLaVA-1.5 training protocol, to investigate the importance of visua l representations for MLLM's spatial intelligence. Furthermore, we create Spatia lBench, a benchmark specifically designed to evaluate CLIP and MLLM in spatial r easoning. SpatialCLIP and SpatialLLaVA achieve substantial performance improveme nts, demonstrating stronger capabilities in spatial perception and reasoning, wh ile maintaining comparable results on general-purpose benchmarks.

****

Mono2Stereo: A Benchmark and Empirical Study for Stereo Conversion

Songsong Yu, Yuxin Chen, Zhongang Qi, Zeke Xie, Yifan Wang, Lijun Wang, Ying Sha n, Huchuan Lu; Proceedings of the Computer Vision and Pattern Recognition Confer ence (CVPR), 2025, pp. 21847-21856

With the rapid proliferation of 3D devices and the shortage of 3D content, stere o conversion is attracting increasing attention. Recent works introduce pretrain ed Diffusion Models (DMs) into this task. However, due to the scarcity of large-scale training data and comprehensive benchmarks, the optimal methodologies for employing DMs in stereo conversion and the accurate evaluation of stereo effects remain largely unexplored. In this work, we introduce the Mono2Stereo dataset, providing high-quality training data and benchmark to support in-depth explorati on of stereo conversion. With this dataset, we conduct an empirical study that y ields two primary findings. 1) The differences between the left and right views are subtle, yet existing metrics consider overall pixels, failing to concentrate on regions critical to stereo effects. 2) Mainstream methods adopt either one-s tage left-to-right generation or warp-and-inpaint pipeline, facing challenges of degraded stereo effect and image distortion respectively. Based on these findin gs, we introduce a new evaluation metric, Stereo Intersection-over-Union, which

prioritizes disparity and achieves a high correlation with human judgments on st
ereo effect. Moreover, we propose a strong baseline model, harmonizing the stere
o effect and image quality simultaneously, and notably surpassing current mainst
ream methods. Our code and data will be open-sourced to promote further research
 in stereo conversion.
****

Towards Open-Vocabulary Audio-Visual Event Localization
Jinxing Zhou, Dan Guo, Ruohao Guo, Yuxin Mao, Jingjing Hu, Yiran Zhong, Xiaojun
Chang, Meng Wang; Proceedings of the Computer Vision and Pattern Recognition Con
ference (CVPR), 2025, pp. 8362-8371
The Audio-Visual Event Localization (AVEL) task aims to temporally locate and cl
assify video events that are both audible and visible.Most research in this fiel
d assumes a closed-set setting, which restricts these models' ability to handle
test data containing event categories absent (unseen) during training. Recently,
 a few studies have explored AVEL in an open-set setting, enabling the recogniti
on of unseen events as "unknown", but without providing category-specific semant
ics.In this paper, we advance the field by introducing the Open-Vocabulary Audio
-Visual Event Localization (OV-AVEL) problem, which requires localizing audio-vi
sual events and predicting explicit categories for both seen and unseen data at
inference.To address this new task, we propose the OV-AVEBench dataset, comprisi
ng 24,800 videos across 67 real-life audio-visual scenes (seen:unseen = 46:21),
each with manual segment-level annotation.We also establish three evaluation met
rics for this task.Moreover, we investigate two baseline approaches, one trainin
g-free and one using a further fine-tuning paradigm.Specifically, we utilize the
 unified multimodal space from the pretrained ImageBind model to extract audio,
visual, and textual (event classes) features.The training-free baseline then det
ermines predictions by comparing the consistency of audio-text and visual-text f
eature similarities.The fine-tuning baseline incorporates lightweight temporal l
ayers to encode temporal relations within the audio and visual modalities, using
 OV-AVEBench training data for model fine-tuning.We evaluate these baselines on
the proposed OV-AVEBench dataset and discuss potential directions for future wor
k in this new field.
****

One-shot 3D Object Canonicalization based on Geometric and Semantic Consistency
Li Jin, Yujie Wang, Wenzheng Chen, Qiyu Dai, Qingzhe Gao, Xueying Qin, Baoquan C
hen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR
), 2025, pp. 16850-16859
3D object canonicalization is a fundamental task, essential for various downstre
am tasks. Existing methods rely on either cumbersome manual processes or priors
learned from extensive, per-category training samples. Real-world datasets, howe
ver, often exhibit long-tail distributions, challenging existing learning-based
methods, especially in categories with limited samples. We address this by intro
ducing the first one-shot category-level object canonicalization framework that
operates under arbitrary poses, requiring only a single canonical model as a ref
erence (the "prior model") for each category. To canonicalize any object, our fr
amework first extracts semantic cues with large language models (LLMs) and visio
n-language models (VLMs) to establish correspondences with the prior model. We i
ntroduce a novel joint energy function to enforce geometric and semantic consist
ency, aligning object orientations precisely despite significant shape variation
s. Moreover, we adopt a support-plane strategy to reduce search space for initia
l poses and utilize a semantic relationship map to select the canonical pose fro
m multiple hypotheses. Extensive experiments on multiple datasets demonstrate th
at our framework achieves state-of-the-art performance and validates key design
choices. Using our framework, we create the Canonical Objaverse Dataset (COD), c
anonicalizing 32K samples in the Objaverse-LVIS dataset, underscoring the effect
iveness of our framework on handling large-scale datasets. Project page at https
://jinli998.github.io/One-shot_3D_Object_Canonicalization/
****

Inst3D-LMM: Instance-Aware 3D Scene Understanding with Multi-modal Instruction T
uning

Hanxun Yu, Wentong Li, Song Wang, Junbo Chen, Jianke Zhu; Proceedings of the Com puter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14147-14157
Despite encouraging progress in 3D scene understanding, it remains challenging t o develop an effective Large Multi-modal Model (LMM) that is capable of understa nding and reasoning in complex 3D environments. Most previous methods typically encode 3D point and 2D image features separately, neglecting interactions betwee n 2D semantics and 3D object properties, as well as the spatial relationships wi thin the 3D environment. This limitation not only hinders comprehensive represen tations of 3D scene, but also compromises training and inference efficiency. To address these challenges, we propose a unified Instance-aware 3D Large Multi-mod al Model (Inst3D-LMM) to deal with multiple 3D scene understanding tasks simulta neously. To obtain the fine-grained instance-level visual tokens, we first intro duce a novel Multi-view Cross-Modal Fusion (MCMF) module to inject the multi-vie w 2D semantics into their corresponding 3D geometric features. For scene-level r elation-aware tokens, we further present a 3D Instance Spatial Relation (3D-ISR) module to capture the intricate pairwise spatial relationships among objects. A dditionally, we perform end-to-end multi-task instruction tuning simultaneously without the subsequent task-specific fine-tuning. Extensive experiments demonstr ate that our approach outperforms the state-of-the-art methods across 3D scene u nderstanding, reasoning and grounding tasks. Source code is available at: https: //github.com/hanxunyu/Inst3D-LMM.
****

S2Gaussian: Sparse-View Super-Resolution 3D Gaussian Splatting
Yecong Wan, Mingwen Shao, Yuanshuo Cheng, Wangmeng Zuo; Proceedings of the Compu ter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 711-721
In this paper, we aim ambitiously for a realistic yet challenging problem, namel y, how to reconstruct high-quality 3D scenes from sparse low-resolution views th at simultaneously suffer from deficient perspectives and contents. Whereas exist ing methods only deal with either sparse views or low-resolution observations, t hey fail to handle such hybrid and complicated scenarios. To this end, we propos e a novel Sparse-view Super-resolution 3D Gaussian Splatting framework, dubbed S 2Gaussian, that can reconstruct structure-accurate and detail-faithful 3D scenes with only sparse and low-resolution views. The S2Gaussian operates in a two-sta ge fashion. In the first stage, we initially optimize a low-resolution Gaussian representation with depth regularization and densify it to initialize the high-r esolution Gaussians through a tailored Gaussian Shuffle operation. In the second stage, we refine the high-resolution Gaussians with the super-resolved images g enerated from both original sparse views and pseudo-views rendered by the low-re solution Gaussians. In which a customized blur-free inconsistency modeling schem e and a 3D robust optimization strategy are elaborately designed to mitigate mul ti-view inconsistency and eliminate erroneous updates caused by imperfect superv ision. Extensive experiments demonstrate superior results and in particular esta blishing new state-of-the-art performances with more consistent geometry and fin er details. Project Page https://jeasco.github.io/S2Gaussian/.
****

HIIF: Hierarchical Encoding based Implicit Image Function for Continuous Super-r esolution
Yuxuan Jiang, Ho Man Kwan, Tianhao Peng, Ge Gao, Fan Zhang, Xiaoqing Zhu, Joel S ole, David Bull; Proceedings of the Computer Vision and Pattern Recognition Conf erence (CVPR), 2025, pp. 2289-2299
Recent advances in implicit neural representations (INRs) have shown significant promise in modeling visual signals for various low-vision tasks including image super-resolution (ISR). INR-based ISR methods typically learn continuous repres entations, providing flexibility for generating high-resolution images at any de sired scale from their low-resolution counterparts. However, existing INR-based ISR methods utilize multi-layer perceptrons for parameterization in the network; this does not take account of the hierarchical structure existing in local samp ling points and hence constrains the representation capability. In this paper, w e propose a new Hierarchical encoding based Implicit Image Function for continuo us image super-resolution, HIIF, which leverages a novel hierarchical positional

encoding that enhances the local implicit representation, enabling it to capture fine details at multiple scales. Our approach also embeds a multi-head linear attention mechanism within the implicit attention network by taking additional non-local information into account. Our experiments show that, when integrated with different backbone encoders, HIIF outperforms the state-of-the-art continuous image super-resolution methods by up to 0.17dB in PSNR. The source code of HIIF will be made publicly available at www.github.com.

****

Motion Prompting: Controlling Video Generation with Motion Trajectories

Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, Chen Sun, Oliver Wang, Andrew Owens, Deqing Sun; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1-12

Motion control is crucial for generating expressive and compelling video content; however, most existing video generation models rely mainly on text prompts for control, which struggle to capture the nuances of dynamic actions and temporal compositions. To this end, we train a video generation model conditioned on spatio-temporally sparse or dense motion trajectories. In contrast to prior motion conditioning work, this flexible representation can encode any number of trajectories, object-specific or global scene motion, and temporally sparse motion; due to its flexibility we refer to this conditioning as motion prompts. While users may directly specify sparse trajectories, we also show how to translate high-level user requests into detailed, semi-dense motion prompts, a process we term motion prompt expansion. We demonstrate the versatility of our approach through various applications, including camera and object motion control, "interacting" with an image, motion transfer, and image editing. Our results showcase emergent behaviors, such as realistic physics, suggesting the potential of motion prompts for probing video models and interacting with future generative world models. Finally, we evaluate quantitatively, conduct a human study, and demonstrate strong performance.

****

VERA: Explainable Video Anomaly Detection via Verbalized Learning of Vision-Language Models

Muchao Ye, Weiyang Liu, Pan He; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8679-8688

The rapid advancement of vision-language models (VLMs) has established a new paradigm in video anomaly detection (VAD): leveraging VLMs to simultaneously detect anomalies and provide comprehendible explanations for the decisions. Existing work in this direction often assumes the complex reasoning required for VAD exceeds the capabilities of pretrained VLMs. Consequently, these approaches either incorporate specialized reasoning modules during inference or rely on instruction tuning datasets through additional training to adapt VLMs for VAD. However, such strategies often incur substantial computational costs or data annotation overhead. To address these challenges in explainable VAD, we introduce a verbalized learning framework named VERA that enables VLMs to perform VAD without model parameter modifications. Specifically, VERA automatically decomposes the complex reasoning required for VAD into reflections on simpler, more focused guiding questions capturing distinct abnormal patterns. It treats these reflective questions as learnable parameters and optimizes them through data-driven verbal interactions between learner and optimizer VLMs, using coarsely labeled training data. During inference, VERA embeds the learned questions into model prompts to guide VLMs in generating segment-level anomaly scores, which are then refined into frame-level scores via the fusion of scene and temporal contexts. Experimental results on challenging benchmarks demonstrate that the learned questions of VERA are highly adaptable, significantly improving both detection performance and explainability of VLMs for VAD.

****

ProHOC: Probabilistic Hierarchical Out-of-Distribution Classification via Multi-Depth Networks

Erik Wallin, Fredrik Kahl, Lars Hammarstrand; Proceedings of the Computer Vision

and Pattern Recognition Conference (CVPR), 2025, pp. 20612-20621
Out-of-distribution (OOD) detection in deep learning has traditionally been framed as a binary task, where samples are either classified as belonging to the known classes or marked as OOD, with little attention given to the semantic relationships between OOD samples and the in-distribution (ID) classes. We propose a framework for detecting and classifying OOD samples in a given class hierarchy. Specifically, we aim to predict OOD data to their correct internal nodes of the class hierarchy, whereas the known ID classes should be predicted as their corresponding leaf nodes. Our approach leverages the class hierarchy to create a probabilistic model and we implement this model by using networks trained for ID classification at multiple hierarchy depths. We conduct experiments on three datasets with predefined class hierarchies and show the effectiveness of our method. Our code is available at https://github.com/walline/prohoc.
****

CCIN: Compositional Conflict Identification and Neutralization for Composed Image Retrieval
Likai Tian, Jian Zhao, Zechao Hu, Zhengwei Yang, Hao Li, Lei Jin, Zheng Wang, Xuelong Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3974-3983
Composed Image Retrieval (CIR) is a multi-modal task that seeks to retrieve target images by harmonizing a reference image with a modified instruction. A key challenge in CIR lies in compositional conflicts between the reference image (e.g., blue, long sleeve) and the modified instruction (e.g., grey, short sleeve). Previous works attempt to mitigate such conflicts through feature-level manipulation, commonly employing learnable masks to obscure conflicting features within the reference image. However, the inherent complexity of feature spaces poses significant challenges in precise conflict neutralization, thereby leading to uncontrollable results. To this end, this paper proposes the Compositional Conflict Identification and Neutralization (CCIN) framework, which sequentially identifies and neutralizes compositional conflicts for effective CIR. Specifically, CCIN comprises two core modules: 1) Compositional Conflict Identification module, which utilizes LLM-based analysis to identify specific conflicting attributes, and 2) Compositional Conflict Neutralization module, which first generates a kept instruction to preserve non-conflicting attributes, then neutralizes conflicts under collaborative guidance of both the kept and modified instructions. Extensive experiments demonstrate the superiority of CCIN over the state-of-the-arts. Code repository: https://github.com/LikaiTian/CCIN.
****

CLIP is Strong Enough to Fight Back: Test-time Counterattacks towards Zero-shot Adversarial Robustness of CLIP
Songlong Xing, Zhengyu Zhao, Nicu Sebe; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15172-15182
Despite its prevalent use in image-text matching tasks in a zero-shot manner, CLIP has been shown to be highly vulnerable to adversarial perturbations added onto images. Recent studies propose to finetune the vision encoder of CLIP with adversarial samples generated on the fly, and show improved robustness against adversarial attacks on a spectrum of downstream datasets, a property termed as zero-shot robustness. In this paper, we show that malicious perturbations that seek to maximise the classification loss lead to `falsely stable' images, and propose to leverage the pre-trained vision encoder of CLIP to counterattack such adversarial images during inference to achieve robustness. Our paradigm is simple and training-free, providing the first method to defend CLIP from adversarial attacks at test time, which is orthogonal to existing methods aiming to boost zero-shot adversarial robustness of CLIP. We conduct experiments across 16 classification datasets, and demonstrate stable and consistent gains compared to test-time defence methods adapted from existing adversarial robustness studies that do not rely on external networks, without noticeably impairing performance on clean images. We also show that our paradigm can be employed on CLIP models that have been adversarially finetuned to further enhance their robustness at test time. Our code will be released.

****
OverLoCK: An Overview-first-Look-Closely-next ConvNet with Context-Mixing Dynamic Kernels
Meng Lou, Yizhou Yu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 128-138
Top-down attention plays a crucial role in the human vision system, wherein the brain initially obtains a rough overview of a scene to discover salient cues (i.e., overview first), followed by a more careful finer-grained examination (i.e., look closely next). However, modern ConvNets remain confined to a pyramid structure that successively downsamples the feature map for receptive field expansion, neglecting this crucial biomimetic principle. We present OverLoCK, the first pure ConvNet backbone architecture that explicitly incorporates a top-down attention mechanism. Unlike pyramid backbone networks, our design features a branched architecture with three synergistic sub-networks: 1) a Base-Net that encodes low/mid-level features; 2) a lightweight Overview-Net that generates dynamic top-down attention through coarse global context modeling (i.e., overview first); and 3) a robust Focus-Net that performs finer-grained perception guided by top-down attention (i.e., look closely next). To fully unleash the power of top-down attention, we further propose a novel context-mixing dynamic convolution (ContMix) that effectively models long-range dependencies while preserving inherent local inductive biases even when the input resolution increases, addressing critical limitations in existing convolutions. Our OverLoCK exhibits a notable performance improvement over existing methods. For instance, OverLoCK-T achieves a Top-1 accuracy of 84.2%, significantly surpassing ConvNeXt-B while using only around one-third of the FLOPs/parameters. On object detection, our OverLoCK-S clearly surpasses MogaNet-B by 1% in AP^b. On semantic segmentation, our OverLoCK-T remarkably improves UniRepLKNet-T by 1.7% in mIoU. Code is publicly available at https://rb.gy/wit4jh.
****
SoftShadow: Leveraging Soft Masks for Penumbra-Aware Shadow Removal
Xinrui Wang, Lanqing Guo, Xiyu Wang, Siyu Huang, Bihan Wen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23217-23226
Recent advancements in deep learning have yielded promising results for the image shadow removal task. However, most existing methods rely on binary pre-generated shadow masks. The binary nature of such masks could potentially lead to artifacts near the boundary between shadow and non-shadow areas. In view of this, inspired by the physical model of shadow formation, we introduce novel soft shadow masks specifically designed for shadow removal. To achieve such soft masks, we propose a SoftShadow framework by leveraging the prior knowledge of pretrained SAM and integrating physical constraints. Specifically, we jointly tune the SAM and the subsequent shadow removal network using penumbra formation constraint loss, mask reconstruction loss, and shadow removal loss. This framework enables accurate predictions of penumbra (partially shaded) and umbra (fully shaded) areas while simultaneously facilitating end-to-end shadow removal. Through extensive experiments on popular datasets, we found that our SoftShadow framework, which generates soft masks, can better restore boundary artifacts, achieve state-of-the-art performance, and demonstrate superior generalizability.
****
Graph-Embedded Structure-Aware Perceptual Hashing for Neural Network Protection and Piracy Detection
Ruiheng Liu, Haozhe Chen, Boyao Zhao, Kejiang Chen, Weiming Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20169-20178
The advancement of AI technology has significantly influenced production activities, increasing the focus on copyright protection for AI models. Model perceptual hashing offers an efficient solution for retrieving the pirated models. Existing methods, such as handcrafted feature-based and dual-branch network-based perceptual hashing, have proven effective in detecting pirated models. However, these approaches often struggle to differentiate non-pirated models, leading to frequent false positives in model authentication and protection. To address this cha

llenge, this paper proposes a structurally-aware perceptual model hashing techni que that achieved reduced false positives while maintaining high true positive r ates. Specifically, we introduce a method for converting the diverse neural netw ork structures into graph structures suitable for DNN processing, then utilize a graph neural network to learn their structural features representation. Our app roach integrates perceptual parameter-based model hashing, achieving robust perf ormance with higher detection accuracy and fewer false positives. Experimental r esults show that the proposed method has only 3% false positive rate when detect ing the non-pirated model, and the detection accuracy of pirated model reaches m ore than 98%.
****

VTON-HandFit: Virtual Try-on for Arbitrary Hand Pose Guided by Hand Priors Embed ding
Yujie Liang, Xiaobin Hu, Boyuan Jiang, Donghao Luo, Xu Peng, Kai Wu, Chengming X u, Wenhui Han, Taisong Jin, Chengjie Wang, Rongrong Ji; Proceedings of the Compu ter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22616-22626
Although diffusion-based image virtual try-on has made considerable progress, em erging approaches still struggle to effectively address the issue of hand occlus ion (i.e., clothing regions occluded by the hand part), leading to a notable deg radation of the try-on performance. To tackle this issue widely existing in real -world scenarios, we propose VTON-HandFit, leveraging the power of hand priors t o reconstruct the appearance and structure for hand occlusion cases. Firstly, we tailor a Hand-Pose Aggregation Net using the ControlNet-based structure explici tly and adaptively encoding the global hand and pose priors. Besides, to fully e xploit the hand-related structure and appearance information, we propose Hand-fe ature Disentanglement Embedding module to disentangle the hand priors into the h and structure-parametric and visual-appearance features, and customize a masked cross attention for further decoupled feature embedding. Lastly, we customize a hand-canny constraint loss to better learn the structure edge knowledge from the hand template of model image. VTON-HandFit outperforms the baselines in qualita tive and quantitative evaluations on the public dataset and our self-collected h and-occlusion Handfit-3K dataset particularly for the arbitrary hand pose occlus ion cases in real-world scenarios.
****

Interleaved-Modal Chain-of-Thought
Jun Gao, Yongqi Li, Ziqiang Cao, Wenjie Li; Proceedings of the Computer Vision a nd Pattern Recognition Conference (CVPR), 2025, pp. 19520-19529
Chain-of-Thought (CoT) prompting elicits large language models (LLMs) to produce a series of intermediate reasoning steps before arriving at the final answer.Ho wever, when transitioning to vision-language models (VLMs), their text-only rati onales struggle to express the fine-grained associations with the original image .In this paper, we propose an image-incorporated multimodal Chain-of-Thought, na med Interleaved-modal Chain-of-Thought (ICoT), which generates sequential reason ing steps consisting of paired visual and textual rationales to infer the final answer.Intuitively, the novel ICoT requires VLMs to enable the generation of fin e-grained interleaved-modal content, which is hard for current VLMs to fulfill.C onsidering that the required visual information is usually part of the input ima ge, we propose Attention-driven Selection (ADS) to realize ICoT over existing VL Ms.ADS intelligently inserts regions of the input image to generate the interlea ved-modal reasoning steps with ignorable additional latency.ADS relies solely on the attention map of VLMs without the need for parameterization, and therefore it is a plug-and-play strategy that can be generalized to a spectrum of VLMs.We apply ADS to realize ICoT on two popular VLMs of different architectures.Extensi ve evaluations of three benchmarks have shown that ICoT prompting achieves subst antial performance (up to 14%) and interpretability improvements compared to exi sting multimodal CoT prompting methods.
****

Uni-Renderer: Unifying Rendering and Inverse Rendering Via Dual Stream Diffusion
Zhifei Chen, Tianshuo Xu, Wenhang Ge, Leyi Wu, Dongyu Yan, Jing He, Luozhou Wang , Lu Zeng, Shunsi Zhang, Ying-Cong Chen; Proceedings of the Computer Vision and

Pattern Recognition Conference (CVPR), 2025, pp. 26504-26513
Rendering and inverse rendering are pivotal tasks in both computer vision and gr
aphics. The rendering equation is the core of the two tasks, as an ideal conditi
onal distribution transfer function from intrinsic properties to RGB images. Des
pite achieving promising results of existing rendering methods, they merely appr
oximate the ideal estimation for a specific scene and come with a high computati
onal cost. Additionally, the inverse conditional distribution transfer is intrac
table due to the inherent ambiguity. To address these challenges, we propose a d
ata-driven method that jointly models rendering and inverse rendering as two con
ditional generation tasks within a single diffusion framework. Inspired by UniDi
ffuser, we utilize two distinct time schedules to model both tasks, and with a t
ailored dual streaming module, we achieve cross-conditioning of two pre-trained
diffusion models. This unified approach, named Uni-Renderer, allows the two proc
esses to facilitate each other through a cycle-consistent constrain, mitigating
ambiguity by enforcing consistency between intrinsic properties and rendered ima
ges.Combined with a meticulously prepared dataset, our method effectively decomp
osition of intrinsic properties and demonstrating a strong capability to recogni
ze changes during rendering. We will open-source our training and inference code
to the public, fostering further research and development in this area.
****

Enhancing Adversarial Transferability with Checkpoints of a Single Model's Train
ing
Shixin Li, Chaoxiang He, Xiaojing Ma, Bin Benjamin Zhu, Shuo Wang, Hongsheng Hu,
Dongmei Zhang, Linchen Yu; Proceedings of the Computer Vision and Pattern Recog
nition Conference (CVPR), 2025, pp. 20685-20694
Adversarial attacks threaten the integrity of deep neural networks (DNNs), parti
cularly in high-stakes applications. In this paper, we present a novel black-box
adversarial attack that leverages the diverse checkpoints generated during a si
ngle model's training trajectory. Unlike conventional ensemble attacks that requ
ire multiple surrogate models with diverse architectures, our approach exploits
the intrinsic diversity captured over different training stages of a single surr
ogate model. By decomposing the learned representations into task-intrinsic and
task-irrelevant components, we employ an accuracy gap-based selection strategy t
o identify checkpoints that predominantly capture transferable, task-intrinsic k
nowledge. Extensive experiments on ImageNet and CIFAR-10 demonstrate that our me
thod consistently outperforms traditional ensemble attacks in terms of transfera
bility, even under resource-constrained and practical settings. This work offers
a resource-efficient solution for crafting highly transferable adversarial exam
ples and provides new insights into the dynamics of adversarial vulnerability.
****

POSTA: A Go-to Framework for Customized Artistic Poster Generation
Haoyu Chen, Xiaojie Xu, Wenbo Li, Jingjing Ren, Tian Ye, Songhua Liu, Ying-Cong
Chen, Lei Zhu, Xinchao Wang; Proceedings of the Computer Vision and Pattern Reco
gnition Conference (CVPR), 2025, pp. 28694-28704
Poster design is a critical medium for visual communication. Prior work has expl
ored automatic poster design using deep learning techniques, but these approache
s lack text accuracy, user customization, and aesthetic appeal, limiting their a
pplicability in artistic domains such as movies and exhibitions, where both clea
r content delivery and visual impact are essential. To address these limitations
, we present POSTA: a modular framework powered by diffusion models and multimod
al large language models (MLLMs) for customized artistic poster generation. The
framework consists of three modules. Background Diffusion creates a themed backg
round based on user input. Design MLLM then generates layout and typography elem
ents that align with and complement the background style. Finally, to enhance th
e poster's aesthetic appeal, ArtText Diffusion applies additional stylization to
key text elements. The final result is a visually cohesive and appealing poster
, with a fully modular process that allows for complete customization. To train
our models, we develop the PosterArt dataset, comprising high-quality artistic p
osters annotated with layout, typography, and pixel-level stylized text segmenta
tion. Our comprehensive experimental analysis demonstrates POSTA's exceptional c

ontrollability and design diversity, outperforming existing models in both text accuracy and aesthetic quality.
****

NSD-Imagery: A Benchmark Dataset for Extending fMRI Vision Decoding Methods to Mental Imagery

Reese Kneeland, Paul S. Scotti, Ghislain St-Yves, Jesse Breedlove, Kendrick Kay, Thomas Naselaris; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28852-28862

We release NSD-Imagery, a benchmark dataset of human fMRI activity paired with mental images, to complement the existing Natural Scenes Dataset (NSD), a large-scale dataset of fMRI activity paired with seen images that enabled unprecedented improvements in fMRI-to-image reconstruction efforts. Recent models trained on NSD have been evaluated only on seen image reconstruction. Using NSD-Imagery, it is possible to assess how well these models perform on mental image reconstruction. This is a challenging generalization requirement because mental images are encoded in human brain activity with relatively lower signal-to-noise and spatial resolution; however, generalization from seen to mental imagery is critical for real-world applications in medical domains and brain-computer interfaces, where the desired information is always internally generated. We provide benchmarks for a suite of recent NSD-trained open-source visual decoding models (MindEye1, MindEye2, Brain Diffuser, iCNN, Takagi et al.) on NSD-Imagery, and show that the performance of decoding methods on mental images is largely decoupled from performance on vision reconstruction. We further demonstrate that architectural choices significantly impact cross-decoding performance: models employing simple linear decoding architectures and multimodal feature decoding generalize better to mental imagery, while complex architectures tend to overfit visual training data. Our findings indicate that mental imagery datasets are critical for the development of practical applications, and establish NSD-Imagery as a useful resource for better aligning visual decoding methods with this goal.
****

VLsI: Verbalized Layers-to-Interactions from Large to Small Vision Language Models

Byung-Kwan Lee, Ryo Hachiuma, Yu-Chiang Frank Wang, Yong Man Ro, Yueh-Hua Wu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29545-29557

The recent surge in high-quality visual instruction tuning samples from closed-source vision-language models (VLMs) such as GPT-4V has accelerated the release of open-source VLMs across various model sizes. However, scaling VLMs to improve performance using larger models brings significant computational challenges, especially for deployment on resource-constrained devices like mobile platforms and robots. To address this, we propose VLsI: Verbalized Layers-to-Interactions, a new VLM family in 2B and 7B model sizes, which prioritizes efficiency without compromising accuracy. VLsI leverages a unique, layer-wise distillation process, introducing intermediate "verbalizers" that map features from each layer to natural language space, allowing smaller VLMs to flexibly align with the reasoning processes of larger VLMs. This approach mitigates the training instability often encountered in output imitation and goes beyond typical final-layer tuning by aligning the small VLMs' layer-wise progression with that of the large ones. We validate VLsI across ten challenging vision-language benchmarks, achieving notable performance gains (11.0% for 2B and 17.4% for 7B) over GPT-4V without the need for model scaling, merging, or architectural changes.
****

O-TPT: Orthogonality Constraints for Calibrating Test-time Prompt Tuning in Vision-Language Models

Ashshak Sharifdeen, Muhammad Akhtar Munir, Sanoojan Baliah, Salman Khan, Muhammad Haris Khan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19942-19951

Test-time prompt tuning for vision-language models (VLMs) is getting attention because of their ability to learn with unlabeled data without fine-tuning. Although test-time prompt tuning methods for VLMs can boost accuracy, the resulting mo

dels tend to demonstrate poor calibration, which casts doubts on the reliability and trustworthiness of these models. Notably, more attention needs to be devoted to calibrating the test-time prompt tuning in vision-language models. To this end, we propose a new approach, called O-TPT that introduces orthogonality constraints on the textual features corresponding to the learnable prompts for calibrating test-time prompt tuning in VLMs. Towards introducing orthogonality constraints, we make the following contributions. First, we uncover new insights behind the suboptimal calibration performance of existing methods relying on textual feature dispersion. Second, we show that imposing a simple orthogonalization of textual features is a more effective approach towards obtaining textual dispersion. We conduct extensive experiments on various datasets with different backbones and baselines. The results indicate that our method consistently outperforms the prior state of the art in significantly reducing the overall average calibration error. Also, our method surpasses the zero-shot calibration performance on fine-grained classification tasks. Our code is available at https://github.com/ashshaksharifdeen/O-TPT.

****

Just Dance with pi! A Poly-modal Inductor for Weakly-supervised Video Anomaly Detection

Snehashis Majhi, Giacomo D'Amicantonio, Antitza Dantcheva, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, Egor Bondarev, Francois Bremond; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24265-24274

Weakly-supervised methods for video anomaly detection (VAD) are conventionally based merely on RGB spatio-temporal features, which continues to limit their reliability in real-world scenarios. This is due to the fact that RGB-features are not sufficiently distinctive in setting apart categories such as shoplifting from visually similar events. Therefore, towards robust complex real-world VAD, it is essential to augment RGB spatio-temporal features by additional modalities. Motivated by this, we introduce the Poly-modal Induced framework for VAD: PI-VAD (or \pi-VAD), a novel approach that augments RGB representations by five additional modalities. Specifically, the modalities include sensitivity to fine-grained motion (Pose), three dimensional scene and entity representation (Depth), surrounding objects (Panoptic masks), global motion (optical flow), as well as language cues (VLM). Each modality represents an axis of a polygon, streamlined to add salient cues to RGB. \pi-VAD includes two plug-in modules, namely Pseudo-modality Generation module and Cross Modal Induction module, which generate modality-specific prototypical representation and, thereby, induce multi-modal information into RGB cues. These modules operate by performing anomaly-aware auxiliary tasks and necessitate five modality backbones -- only during training. Notably, \pi-VAD achieves state-of-the-art accuracy on three prominent VAD datasets encompassing real-world scenarios, without requiring the computational overhead of five modality backbones at inference.

****

Flash3D: Super-scaling Point Transformers through Joint Hardware-Geometry Locality

Liyan Chen, Gregory P. Meyer, Zaiwei Zhang, Eric M. Wolff, Paul Vernaza; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6595-6604

Recent efforts recognize the power of scale in 3D learning (e.g. PTv3) and attention mechanisms (e.g. FlashAttention).However, current point cloud backbones fail to holistically unify geometric locality, attention mechanisms, and GPU architectures in one view.In this paper, we introduce Flash3D Transformer, which aligns geometric locality and GPU tiling through a principled locality mechanism based on Perfect Spatial Hashing (PSH).The common alignment with GPU tiling naturally fuses our PSH locality mechanism with FlashAttention at negligible extra cost. This mechanism affords flexible design choices throughout the backbone that result in superior downstream task results.Flash3D outperforms state-of-the-art PTv3 results on benchmark datasets, delivering a 2.25x speed increase and 2.4x memory efficiency boost. This efficiency enables scaling to wider attention scopes an

d larger models without additional overhead.Such scaling allows Flash3D to achie
ve even higher task accuracies than PTv3 under the same compute budget.
****

Analyzing the Synthetic-to-Real Domain Gap in 3D Hand Pose Estimation
Zhuoran Zhao, Linlin Yang, Pengzhan Sun, Pan Hui, Angela Yao; Proceedings of the
 Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12255-1226
5
Recent synthetic 3D human datasets for the face, body, and hands have pushed the
 limits on photorealism. Face recognition and body pose estimation have achieved
 state-of-the-art performance using synthetic training data alone, but for the h
and, there is still a large synthetic-to-real gap. This paper presents the first
 systematic study of the synthetic-to-real gap of 3D hand pose estimation. We an
alyze the gap and identify key components such as the forearm, image frequency s
tatistics, hand pose, and object occlusions. To facilitate our analysis, we prop
ose a data synthesis pipeline to synthesize high-quality data. We demonstrate th
at synthetic hand data can achieve the same level of accuracy as real data when
integrating our identified components, paving the path to use synthetic data alo
ne for hand pose estimation. Code and data are available at: https://github.com/
delaprada/HandSynthesis.git.
****

Feature4X: Bridging Any Monocular Video to 4D Agentic AI with Versatile Gaussian
 Feature Fields
Shijie Zhou, Hui Ren, Yijia Weng, Shuwang Zhang, Zhen Wang, Dejia Xu, Zhiwen Fan
, Suya You, Zhangyang Wang, Leonidas Guibas, Achuta Kadambi; Proceedings of the
Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14179-14190
Recent advancements in 2D and multimodal models have achieved remarkable success
 by leveraging large-scale training on extensive datasets. However, extending th
ese achievements to enable free-form interactions and high-level semantic operat
ions with complex 3D/4D scenes remains challenging. This difficulty stems from t
he limited availability of large-scale, annotated 3D/4D or multi-view datasets,
which are crucial for generalizable vision and language tasks such as open-vocab
ulary and prompt-based segmentation, language-guided editing, and visual questio
n answering (VQA). In this paper, we introduce Feature4X, a universal framework
designed to extend any functionality from 2D vision foundation model into the 4D
 realm, using only monocular video input, which is widely available from user-ge
nerated content. The "X" in Feature4X represents its versatility, enabling any t
ask through adaptable, model-conditioned 4D feature field distillation. At the c
ore of our framework is a dynamic optimization strategy that unifies multiple mo
del capabilities into a single representation. Additionally, to the best of our
knowledge, Feature4X is the first method to distill and lift the features of vid
eo foundation models (e.g. SAM2, InternVideo2) into an explicit 4D feature field
 using Gaussian Splatting. Our experiments showcase novel view segment anything,
 geometric and appearance scene editing, and free-form VQA across all time steps
, empowered by LLMs in feedback loops. These advancements broaden the scope of a
gentic AI applications by providing a foundation for scalable, contextually and
spatiotemporally aware systems capable of immersive dynamic 4D scene interaction
.
****

Comprehensive Relighting: Generalizable and Consistent Monocular Human Relightin
g and Harmonization
Junying Wang, Jingyuan Liu, Xin Sun, Krishna Kumar Singh, Zhixin Shu, He Zhang,
Jimei Yang, Nanxuan Zhao, Tuanfeng Y. Wang, Simon S. Chen, Ulrich Neumann, Jae S
hin Yoon; Proceedings of the Computer Vision and Pattern Recognition Conference
(CVPR), 2025, pp. 380-390
This paper introduces Comprehensive Relighting, the first all-in-one approach th
at can both control and harmonize the lighting from an image or video of humans
with arbitrary body parts from any scene. Building such a generalizable model is
 extremely challenging due to the lack of dataset, restricting existing image-ba
sed relighting models to a specific scenario (e.g., face or static human). To ad
dress this challenge, we repurpose a pre-trained diffusion model as a general im

age prior and jointly model the human relighting and background harmonization in the coarse-to-fine framework. To further enhance the temporal coherence of the relighting, we introduce an unsupervised temporal lighting model that learns the lighting cycle consistency from many real-world videos without any ground truth. In inference time, our temporal lighting module is combined with the diffusion models through the spatio-temporal feature blending algorithms without extra training; and we apply a new guided refinement as a post-processing to preserve the high-frequency details from the input image. In the experiments, Comprehensive Relighting shows a strong generalizability and lighting temporal coherence, outperforming existing image-based human relighting and harmonization methods.
****

Hyperspectral Pansharpening via Diffusion Models with Iteratively Zero-Shot Guidance
Jin-Liang Xiao, Ting-Zhu Huang, Liang-Jian Deng, Guang Lin, Zihan Cao, Chao Li, Qibin Zhao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12669-12678
Hyperspectral pansharpening refers to fusing a panchromatic image (PAN) and a low-resolution hyperspectral image (LR-HSI) to obtain a high-resolution hyperspectral image (HR-HSI). Recently, guiding pre-trained diffusion models (DMs) has demonstrated significant potential in this area, leveraging their powerful representational abilities while avoiding complex training processes. However, these DMs are often trained on RGB images, not well-suited for pansharpening tasks, limited in adapting to the hyperspectral images. In this work, we propose a novel guided diffusion scheme with zero-shot guidance and neural spatial-spectral decomposition (NSSD) to iteratively generate the RGB detail image and map the RGB detail image to target HR-HSI. Specifically, zero-shot guidance employs an auxiliary neural network that trained only with a PAN and LR-HSI to guide pre-trained DMs in generating the RGB detail image, informed by specific prior knowledge. Then, NSSD establishes a spectral mapping from the generated RGB detail image to the final HR-HSI. Extensive experiments are conducted on Pavia, Washington DC, Chukusei, and FR1 datasets to demonstrate that the proposed method significantly enhances the performance of DMs for hyperspectral pansharpening tasks, outperforming existing methods across multiple metrics and achieving improvements in visualization results. The code is available at \href https://github.com/Jin-liangXiao/DM-zs https://github.com/Jin-liangXiao/DM-zs .
****

EASEMVC:Efficient Dual Selection Mechanism for Deep Multi-View Clustering
Baili Xiao, Zhibin Dong, Ke Liang, Suyuan Liu, Siwei Wang, Tianrui Liu, Xingchen Hu, En Zhu, Xinwang Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20716-20726
Multi-view clustering represents one of the most established paradigms within the field of unsupervised learning and has witnessed a surge in popularity in recent years. View-pair form contrastive learning allows for consistently representing multiple views by maximizing mutual information between each two views. This approach permits multi-view clustering to discern consistent latent representations across multiple views. However, two significant issues emerge when this approach is considered. i)it is challenging to ascertain which two views are most appropriate for contrastive learning when there are more than three views, particularly without prior knowledge. ii) when all views are included in contrastive learning, multi-view clustering performance is compromised by poor quality views. To tackle these issues, we present a novel Efficient Dual Selection Mechanism for deep Multi-View Clustering framework, termed EASEMVC. Specifically, EASEMVC first constructs a view graph based on the OT distance between the bipartite graph of each view. It then designs a view selection module to realize an efficient view-level selection process through the view topology relations in the view graph structure. Additionally, a cross-view sample graph structure is constructed at the sample level, with the sample topological relations in the cross-view sample graph structure being employed to generate reliable sample learning weights. Based on the view pairs and sample weights selected by the aforementioned method, we employ contrastive learning within the theoretical framework of information

theory to obtain consistent representations between views. Extensive experiments on six benchmark datasets have demonstrated that the proposed EASEMVC outperforms the state-of-the-art methods.
****

Efficient Motion-Aware Video MLLM
Zijia Zhao, Yuqi Huo, Tongtian Yue, Longteng Guo, Haoyu Lu, Bingning Wang, Weipeng Chen, Jing Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24159-24168
Most current video MLLMs rely on uniform frame sampling and image-level encoders, resulting in inefficient data processing and limited motion awareness. To address these challenges, we introduce EMA, an Efficient Motion-Aware video MLLM that utilizes compressed video structures as inputs. We propose a motion-aware GOP (Group of Pictures) encoder that fuses spatial and motion information within a GOP unit in the compressed video stream, generating compact, informative visual tokens. By integrating fewer but denser RGB frames with more but sparser motion vectors in this native slow-fast input architecture, our approach reduces redundancy and enhances motion representation. Additionally, we introduce MotionBench, a benchmark for evaluating motion understanding across four motion types: linear, curved, rotational, and contact-based. Experimental results show that EMA achieves state-of-the-art performance on both MotionBench and popular video question answering benchmarks, while reducing inference costs. Moreover, EMA demonstrates strong scalability, as evidenced by its competitive performance on long video understanding benchmarks.
****

DSPNet: Dual-vision Scene Perception for Robust 3D Question Answering
Jingzhou Luo, Yang Liu, Weixing Chen, Zhen Li, Yaowei Wang, Guanbin Li, Liang Lin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14169-14178
3D Question Answering (3D QA) requires the model to comprehensively understand its situated 3D scene described by the text, then reason about its surrounding environment and answer a question under that situation. However, existing methods usually rely on global scene perception from pure 3D point clouds and overlook the importance of rich local texture details from multi-view images. Moreover, due to the inherent noise in camera poses and complex occlusions, there exists significant feature degradation and reduced feature robustness problems when aligning 3D point cloud with multi-view images. In this paper, we propose a Dual-vision Scene Perception Network (DSPNet), to comprehensively integrate multi-view and point cloud features to improve robustness in 3D QA. Our Text-guided Multi-view Fusion (TGMF) module prioritizes image views that closely match the semantic content of the text. To adaptively fuse back-projected multi-view images with point cloud features, we design the Adaptive Dual-vision Perception (ADVP) module, enhancing 3D scene comprehension. Additionally, our Multimodal Context-guided Reasoning (MCGR) module facilitates robust reasoning by integrating contextual information across visual and linguistic modalities. Experimental results on SQA3D and ScanQA datasets demonstrate the superiority of our DSPNet. Codes will be available at https://github.com/LZ-CH/DSPNet.
****

Zero-Shot 4D Lidar Panoptic Segmentation
Yushan Zhang, Aljoša Ošep, Laura Leal-Taixé, Tim Meinhardt; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24506-24517
Zero-shot 4D segmentation and recognition of arbitrary objects in Lidar is crucial for embodied navigation, with applications ranging from streaming perception to semantic mapping and localization. However, the primary challenge in advancing research and developing generalized, versatile methods for spatio-temporal scene understanding in Lidar lies in the scarcity of datasets that provide the necessary diversity and scale of annotations. To overcome these challenges, we propose SAL-4D (Segment Anything in Lidar--4D), a method that utilizes multi-modal robotic sensor setups as a bridge to distill recent developments in Video Object Segmentation (VOS) in conjunction with off-the-shelf Vision-Language foundation models to Lidar. We utilize VOS models to pseudo-label tracklets in short video

sequences, annotate these tracklets with sequence-level CLIP tokens, and lift th
em to the 4D Lidar space using calibrated multi-modal sensory setups to distill
them to our SAL-4D model. Due to temporal consistent predictions, we outperform
prior art in 3D Zero-Shot Lidar Panoptic Segmentation (LPS) over 5 PQ, and unloc
k Zero-Shot 4D-LPS.
****

MAtCha Gaussians: Atlas of Charts for High-Quality Geometry and Photorealism Fro
m Sparse Views
Antoine Guedon, Tomoki Ichikawa, Kohei Yamashita, Ko Nishino; Proceedings of the
 Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6001-6011
We present a novel appearance model that simultaneously realizes explicit high-q
uality 3D surface mesh recovery and photorealistic novel view synthesis from spa
rse view samples. Our key idea is to model the underlying scene geometry Mesh as
 an Atlas of Charts which we render with 2D Gaussian surfels (MAtCha Gaussians).
 MAtCha distills high-frequency scene surface details from an off-the-shelf mono
cular depth estimator and refines it through Gaussian surfel rendering. The Gaus
sian surfels are attached to the charts on the fly, satisfying photorealism of n
eural volumetric rendering and crisp geometry of a mesh model, i.e., two seeming
ly contradicting goals in a single model. At the core of MAtCha lies a novel neu
ral deformation model and a structure loss that preserve the fine surface detail
s distilled from learned monocular depths while addressing their fundamental sca
le ambiguities. Results of extensive experimental validation demonstrate MAtCha'
s state-of-the-art quality of surface reconstruction and photorealism on-par wit
h top contenders but with dramatic reduction in the number of input views and co
mputational time. We believe MAtCha will serve as a foundational tool for any vi
sual application in vision, graphics, and robotics that require explicit geometr
y in addition to photorealism.
****

Extreme Rotation Estimation in the Wild
Hana Bezalel, Dotan Ankri, Ruojin Cai, Hadar Averbach-Elor; Proceedings of the C
omputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1061-1070
We present a technique and benchmark dataset for estimating the relative 3D orie
ntation between a pair of Internet images captured in an extreme setting, where
the images have limited or non-overlapping field of views. Prior work targeting
extreme rotation estimation assume constrained 3D environments and emulate persp
ective images by cropping regions from panoramic views. However, real images cap
tured in the wild are highly diverse, exhibiting variation in both appearance an
d camera intrinsics. In this work, we propose a Transformer-based method for est
imating relative rotations in extreme real-world settings, and contribute the Ex
tremeLandmarkPairs dataset, assembled from scene-level Internet photo collection
s. Our evaluation demonstrates that our approach succeeds in estimating the rela
tive rotations in a wide variety of extreme-view Internet image pairs, outperfor
ming various baselines, including dedicated rotation estimation techniques and c
ontemporary 3D reconstruction methods.
****

ADU: Adaptive Detection of Unknown Categories in Black-Box Domain Adaptation
Yushan Lai, Guowen Li, Haoyuan Liang, Juepeng Zheng, Zhiyu Ye; Proceedings of th
e Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30588-305
98
Black-box Domain Adaptation (BDA) utilizes a black-box predictor of the source d
omain to label target domain data, addressing privacy concerns in Unsupervised D
omain Adaptation (UDA). However, BDA assumes identical label sets across domain
s, which is unrealistic. To overcome this limitation, we propose a study on BDA
with unknown classes in the target domain. It uses a black-box predictor to labe
l target data and identify "unknown" categories, without requiring access to sou
rce domain data or predictor parameters, thus addressing both data privacy and c
ategory shift issues in traditional UDA. Existing methods face two main challeng
es: (i) Noisy pseudo-labels in knowledge distillation (KD) accumulate prediction
 errors, and (ii) relying on a preset threshold fails to adapt to varying catego
ry shifts. To address these, we propose ADU, a framework that allows the target

domain to autonomously learn pseudo-labels guided by quality and use an adaptive threshold to identify "unknown" categories. Specifically, ADU consists of Selective Amplification Knowledge Distillation (SAKD) and Entopy-Driven Label Differentiation (EDLD). SAKD improves KD by focusing on high-quality pseudo-labels, mitigating the impact of noisy labels. EDLD categorizes pseudo-labels by quality and applies tailored training strategies to distinguish "unknown" categories, improving detection accuracy and adaptability. Extensive experiments show that ADU achieves state-of-the-art results, outperforming the best existing method by 3.1% on VisDA in the OPBDA scenario.
****

EmotiveTalk: Expressive Talking Head Generation through Audio Information Decoupling and Emotional Video Diffusion

Haotian Wang, Yuzhe Weng, Yueyan Li, Zilu Guo, Jun Du, Shutong Niu, Jiefeng Ma, Shan He, Xiaoyan Wu, Qiming Hu, Bing Yin, Cong Liu, Qingfeng Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26212-26221

Diffusion models have revolutionized the field of talking head generation, yet still face challenges in expressiveness, controllability, and stability in long-time generation. In this research, we propose an EmotiveTalk framework to address these issues. Firstly, to realize better control over the generation of lip movement and facial expression, a Vision-guided Audio Information Decoupling (V-AID) approach is designed to generate audio-based decoupled representations aligned with lip movements and expression. Specifically, to achieve alignment between audio and facial expression representation spaces, we present a Diffusion-based Co-speech Temporal Expansion (Di-CTE) module within V-AID to generate expression-related representations under multi-source emotion condition constraints. Then we propose a well-designed Emotional Talking Head Diffusion (ETHD) backbone to efficiently generate highly expressive talking head videos, which contains an Expression Decoupling Injection (EDI) module to automatically decouple the expressions from reference portraits while integrating the target expression information, achieving more expressive generation performance. Experimental results show that EmotiveTalk can generate expressive talking head videos, ensuring the promised controllability of emotions and metric stability during long-time generation, yielding state-of-the-art performance compared to existing methods. The main page of our paper can be found in https://emotivetalk.github.io/.
****

Traversing Distortion-Perception Tradeoff using a Single Score-Based Generative Model

Yuhan Wang, Suzhi Bi, Ying-Jun Angela Zhang, Xiaojun Yuan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2377-2386

The distortion-perception (DP) tradeoff reveals a fundamental conflict between distortion metrics (e.g., MSE and PSNR) and perceptual quality. Recent research has increasingly concentrated on evaluating denoising algorithms within the DP framework. However, existing algorithms either prioritize perceptual quality by sacrificing acceptable distortion, or focus on minimizing MSE for faithful restoration. When the goal shifts or noisy measurements vary, adapting to different points on the DP plane needs retraining or even re-designing the model. Inspired by recent advances in solving inverse problems using score-based generative models, we explore the potential of flexibly and optimally traversing DP tradeoffs using a single pre-trained score-based model. Specifically, we introduce a variance-scaled reverse diffusion process and theoretically characterize the marginal distribution. We then prove that the proposed sample process is an optimal solution to the DP tradeoff for conditional Gaussian distribution. Experimental results on two-dimensional and image datasets illustrate that a single score network can effectively and flexibly traverse the DP tradeoff for general denoising problems.
****

IceDiff: High Resolution and High-Quality Arctic Sea Ice Forecasting with Generative Diffusion Prior

Jingyi Xu, Siwei Tu, Weidong Yang, Ben Fei, Shuhao Li, Keyi Liu, Yeqi Luo, Lipen

g Ma, Lei Bai; Proceedings of the Computer Vision and Pattern Recognition Confer
ence (CVPR), 2025, pp. 10567-10576
Variation of Arctic sea ice has significant impacts on polar ecosystems, transpo
rting routes, coastal communities, and global climate. Tracing the change of sea
 ice at a finer scale is paramount for both operational applications and scienti
fic studies. Recent pan-Arctic sea ice forecasting methods that leverage advance
s in artificial intelligence have made promising progress over numerical models.
 However, forecasting sea ice at higher resolutions is still under-explored. To
bridge the gap, we propose a two-module cooperative deep learning framework, Ice
Diff, to forecast sea ice concentration at finer scales. IceDiff first leverages
 a vision transformer to generate coarse yet superior forecasting results over p
revious methods at a regular 25km grid. This high-quality sea ice forecasting ca
n be utilized as reliable guidance for the next module. Subsequently, an uncondi
tional diffusion model pre-trained on low-resolution sea ice concentration maps
is utilized for sampling down-scaled sea ice forecasting via a zero-shot guided
sampling strategy and a patch-based method. For the first time, IceDiff demonstr
ates sea ice forecasting with a 6.25km resolution. IceDiff extends the boundary
of existing sea ice forecasting models and more importantly, its capability to g
enerate high-resolution sea ice concentration data is vital for pragmatic usages
 and research.
****

DTOS: Dynamic Time Object Sensing with Large Multimodal Model
Jirui Tian, Jinrong Zhang, Shenglan Liu, Luhao Xu, Zhixiong Huang, Gao Huang; Pr
oceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025
, pp. 13810-13820
Existing multimodal large language models (MLLMs) face significant challenges in
 Referring Video Object Segmentation(RVOS). We identify three critical challenge
s: (C1) insufficient quantitative representation of textual numerical data, (C2)
 repetitive and degraded response templates for spatiotemporal referencing, and
(C3) loss of visual information in video sampling queries lacking textual guidan
ce. To address these, we propose a novel framework, Dynamic Time Object Sensing
(DTOS), specifically designed for RVOS. To tackle (C1) and (C2), we introduce sp
ecialized tokens to construct multi-answer response templates, enabling regressi
on of event boundaries and target localization. This approach improves the accur
acy of numerical regression while mitigating the issue of repetitive degradation
. To address (C3), we propose a Text-guided Clip Sampler (TCS) that selects vide
o clips aligned with user instructions, preventing visual information loss and e
nsuring consistent temporal resolution. TCS is also applicable to Moment Retriev
al tasks, with enhanced multimodal input sequences preserving spatial details an
d maximizing temporal resolution. DTOS demonstrates exceptional capability in fl
exibly localizing multiple spatiotemporal targets based on user-provided textual
 instructions. Extensive experiments validate the effectiveness of our approach,
 with DTOS achieving state-of-the-art performance in J&F scores: an improvement
of +4.36 on MeViS, +4.48 on Ref-DAVIS17, and +3.02 on Ref-YT-VOS. Additionally,
our TCS demonstrates exceptional performance in Moment Retrieval. The code is av
ailable at https://github.com/Maulog/OPEN-DTOS-LMM
****

How to Merge Your Multimodal Models Over Time?
Sebastian Dziadzio, Vishaal Udandarao, Karsten Roth, Ameya Prabhu, Zeynep Akata,
 Samuel Albanie, Matthias Bethge; Proceedings of the Computer Vision and Pattern
 Recognition Conference (CVPR), 2025, pp. 20479-20491
Model merging combines expert models---each finetuned from a shared foundation m
odel on diverse tasks and domains---into a single, more capable base model. Howe
ver, existing model merging approaches assume all experts to be available simult
aneously. In reality, new tasks and domains emerge continuously, prompting the n
eed for a dynamic process of integrating these experts over time, which we call
temporal model merging. The temporal dimension introduces unique challenges not
addressed in prior work: At each task, should expert training start from merged
previous experts or the original base model? Should all models be merged at ever
y time step? Which merging techniques are best suited for temporal merging? Shou

ld different strategies be used for the training initialization and deployment p
hases? To tackle these questions, we propose a unified framework called TIME---T
emporal Integration of Model Expertise---that defines temporal model merging acr
oss three axes: (1) Initialization Phase, (2) Deployment Phase, and (3) Merging
Technique. Utilizing TIME, we study temporal model merging across model sizes, t
asks, and compute budgets on the large-scale FoMo-in-Flux benchmark for continua
l multimodal pretraining. Systematic experiments across TIME and FoMo-in-Flux al
low us to arrive at several crucial key insights for temporal model merging to b
etter understand current limits and best practices for successful model merging
across time.
****

Identifying and Mitigating Position Bias of Multi-image Vision-Language Models
Xinyu Tian, Shu Zou, Zhaoyuan Yang, Jing Zhang; Proceedings of the Computer Visi
on and Pattern Recognition Conference (CVPR), 2025, pp. 10599-10609
The evolution of Large Vision-Language Models (LVLMs) has progressed from single
-image understanding to multi-image reasoning. Despite this advancement, our fin
dings indicate that LVLMs struggle to robustly utilize information across multip
le images, with predictions significantly affected by the alteration of image po
sitions. To further explore this issue, we introduce Position-wise Question Answ
ering (PQA), a meticulously designed task to quantify reasoning capabilities at
each position. Our analysis reveals a pronounced position bias in LVLMs: open-so
urce models excel in reasoning with images positioned later but underperform wit
h those in the middle or at the beginning, while proprietary models like GPT-4o
show improved comprehension for images at the beginning and end but struggle wit
h those in the middle. Motivated by these insights, we propose SoFt Attention (S
oFA), a simple, training-free approach that mitigates this bias by employing lin
ear interpolation between inter-image causal attention and bidirectional counter
parts. Experimental results demonstrate that SoFA effectively reduces position b
ias and significantly enhances the reasoning performance of existing LVLMs.
****

Unsupervised Foundation Model-Agnostic Slide-Level Representation Learning
Tim Lenz, Peter Neidlinger, Marta Ligero, Georg Wölflein, Marko van Treeck, Jako
b N. Kather; Proceedings of the Computer Vision and Pattern Recognition Conferen
ce (CVPR), 2025, pp. 30807-30817
Representation learning of pathology whole-slide images (WSIs) has primarily rel
ied on weak supervision with Multiple Instance Learning (MIL). This approach lea
ds to slide representations highly tailored to a specific clinical task. Self-su
pervised learning (SSL) has been successfully applied to train histopathology fo
undation models (FMs) for patch embedding generation. However, generating patien
t or slide level embeddings remains challenging. Existing approaches for slide r
epresentation learning extend the principles of SSL from patch level learning to
 entire slides by aligning different augmentations of the slide or by utilizing
multimodal data. By integrating tile embeddings from multiple FMs, we propose a
new single modality SSL method in feature space that generates useful slide repr
esentations. Our contrastive pretraining strategy, called COBRA, employs multipl
e FMs and an architecture based on Mamba-2. COBRA exceeds performance of state-o
f-the-art slide encoders on four different public Clinical Protemic Tumor Analys
is Consortium (CPTAC) cohorts on average by at least +4.4% AUC, despite only bei
ng pretrained on 3048 WSIs from The Cancer Genome Atlas (TCGA). Additionally, CO
BRA is readily compatible at inference time with previously unseen feature extra
ctors. Code available at https://github.com/KatherLab/COBRA
****

Exploring CLIP's Dense Knowledge for Weakly Supervised Semantic Segmentation
Zhiwei Yang, Yucong Meng, Kexue Fu, Feilong Tang, Shuo Wang, Zhijian Song; Proce
edings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, p
p. 20223-20232
Weakly Supervised Semantic Segmentation (WSSS) with image-level labels aims to a
chieve pixel-level predictions using Class Activation Maps (CAMs). Recently, Con
trastive Language-Image Pre-training (CLIP) has been introduced in WSSS. However
, recent methods primarily focus on image-text alignment for CAM generation, whi

le CLIP's potential in patch-text alignment remains unexplored. In this work, we propose ExCEL to explore CLIP's dense knowledge via a novel patch-text alignment paradigm for WSSS. Specifically, we propose Text Semantic Enrichment (TSE) and Visual Calibration (VC) modules to improve the dense alignment across both text and vision modalities. To make text embeddings semantically informative, our TSE module applies Large Language Models (LLMs) to build a dataset-wide knowledge base and enriches the text representations with an implicit attribute-hunting process. To mine fine-grained knowledge from visual features, our VC module first proposes Static Visual Calibration (SVC) to propagate fine-grained knowledge in a non-parametric manner. Then Learnable Visual Calibration (LVC) is further proposed to dynamically shift the frozen features towards distributions with diverse semantics. With these enhancements, ExCEL not only retains CLIP's training-free advantages but also significantly outperforms other state-of-the-art methods with much less training cost on PASCAL VOC and MS COCO. Code is available at https://github.com/zwyang6/ExCEL.

****

UNIALIGN: Scaling Multimodal Alignment within One Unified Model

Bo Zhou, Liulei Li, Yujia Wang, Huafeng Liu, Yazhou Yao, Wenguan Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29644-29655

We present UNIALIGN, a unified model to align an arbitrary number of modalities (\text e.g. , image, text, audio, 3D point cloud, etc.) through one encoder and a single training phase. Existing solutions typically employ distinct encoders for each modality, resulting in increased parameters as the number of modalities grows. In contrast, UNIALIGN proposes a modality-aware adaptation of the powerful mixture-of-experts (MoE) schema and further integrates it with Low-Rank Adaptation (LoRA), efficiently scaling the encoder to accommodate inputs in diverse modalities while maintaining a fixed computational overhead.Moreover, prior work often requires separate training for each extended modality. This leads to task-specific models and further hinders the communication between modalities.To address this, we propose a soft modality binding strategy that aligns all modalities using unpaired data samples across datasets. Two additional training objectives are introduced to distill knowledge from well-aligned anchor modalities and prior multimodal models, elevating UNIALIGN into a high performance multimodal foundation model.Experiments on 11 benchmarks across 6 different modalities demonstrate that UNIALIGN could achieve comparable performance to SOTA approaches, while using merely 7.8M trainable parameters and maintaining an identical model with the same weight across all tasks. Our code shall be released.

****

ShowHowTo: Generating Scene-Conditioned Step-by-Step Visual Instructions

Tomáš Sou■ek, Prajwal Gatti, Michael Wray, Ivan Laptev, Dima Damen, Josef Sivic; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27435-27445

The goal of this work is to generate step-by-step visual instructions in the form of a sequence of images, given an input image that provides the scene context and the sequence of textual instructions. This is a challenging problem as it requires generating multi-step image sequences to achieve a complex goal while being grounded in a specific environment. Part of the challenge stems from the lack of large-scale training data for this problem. The contribution of this work is thus three-fold. First, we introduce an automatic approach for collecting large step-by-step visual instruction training data from instructional videos. We apply this approach to one million videos and create a large-scale, high-quality dataset of 0.6M sequences of image-text pairs. Second, we develop and train ShowHowTo, a video diffusion model capable of generating step-by-step visual instructions consistent with the provided input image. Third, we evaluate the generated image sequences across three dimensions of accuracy (step, scene, and task) and show our model achieves state-of-the-art results on all of them. Our code, dataset, and trained models are publicly available.

****

Exploration-Driven Generative Interactive Environments

Nedko Savov, Naser Kazemi, Mohammad Mahdi, Danda Pani Paudel, Xi Wang, Luc Van Gool; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27597-27607

Modern world models require costly and time-consuming collection of large video datasets with action demonstrations by people or by environment-specific agents. To simplify training, we focus on using many virtual environments for inexpensive, automatically collected interaction data. Genie, a recent multi-environment world model, demonstrates simulation abilities of many environments with shared behavior. Unfortunately, training their model requires expensive demonstrations. Therefore, we propose a training framework merely using a random agent in virtual environments. While the model trained in this manner exhibits good controls, it is limited by the random exploration possibilities. To address this limitation, we propose AutoExplore Agent - an exploration agent that entirely relies on the uncertainty of the world model, delivering diverse data from which it can learn the best. Our agent is fully independent of environment-specific rewards and thus adapts easily to new environments. With this approach, the pretrained multi-environment model can quickly adapt to new environments achieving video fidelity and controllability improvement. In order to obtain automatically large-scale interaction datasets for pretraining, we group environments with similar behavior and controls. To this end, we annotate the behavior and controls of 974 virtual environments - a dataset that we name RetroAct. For building our model, we first create an open implementation of Genie - GenieRedux and apply enhancements and adaptations in our version GenieRedux-G. Our code and data are available at https://github.com/insait-institute/GenieRedux

****

Task-Agnostic Guided Feature Expansion for Class-Incremental Learning

Bowen Zheng, Da-Wei Zhou, Han-Jia Ye, De-Chuan Zhan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10099-10109

The ability to learn new concepts while preserve the learned knowledge is desirable for learning systems in Class-Incremental Learning (CIL). Recently, feature expansion of the model become a prevalent solution for CIL, where the old features are fixed during the training of the new task while new features are expanded for the new tasks. However, such task-specific features learned from the new task may collide with the old features, leading to misclassification between tasks. Therefore, the expanded model is often encouraged to capture diverse features from the new task, aiming to avoid such collision. However, the existing solution is largely restricted to the samples from the current task, because of the poor accessibility to previous samples. To promote the learning and transferring of diverse features across tasks, we propose a framework called Task-Agnostic Guided Feature Expansion (TagFex). Firstly, it captures task-agnostic features continually with a separate model, providing extra task-agnostic features for subsequent tasks. Secondly, to obtain useful features from the task-agnostic model for the current task, it aggregates the task-agnostic features with the task-specific feature using a merge attention. Then the aggregated feature is transferred back into the task-specific feature for inference, helping the task-specific model capture diverse features. Extensive experiments show the effectiveness and superiority of TagFex on various CIL settings.

****

ShowUI: One Vision-Language-Action Model for GUI Visual Agent

Kevin Qinghong Lin, Linjie Li, Difei Gao, Zhengyuan Yang, Shiwei Wu, Zechen Bai, Stan Weixian Lei, Lijuan Wang, Mike Zheng Shou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19498-19508

Building Graphical User Interface (GUI) assistants holds significant promise for enhancing human workflow productivity. While most agents are language-based, relying on closed-source API with text-rich meta-information (e.g., HTML or accessibility tree), they show limitations in perceiving UI visuals as humans do, highlighting the need for GUI visual agents.In this work, we develop a vision-language-action model in the digital world, namely "Our Model," which features the following innovations:1. **UI-Guided Visual Token Selection** to reduce computational costs by formulating screenshots as a UI-connected graph, adaptively identify

ing their redundant relationships and serving as the criteria for token selection during self-attention blocks. 2. **Interleaved Vision-Language-Action Streaming** that flexibly unifies diverse needs within GUI tasks, enabling effective management of visual-action history in navigation or pairing multi-turn query-action sequences per screenshot to enhance training efficiency. 3. **Small-Scale High-Quality GUI Instruction-Following Datasets** by careful data curation and employing a resampling strategy to address significant data type imbalances. With the above components, our model, a lightweight 2B model using 256K data, achieves a strong 75.1% accuracy in zero-shot screenshot grounding. Its UI-guided token selection further reduces 33% of redundant visual tokens during training and speeds up performance by 1.4x. Navigation experiments across web, mobile, and online environments further underscore the effectiveness and potential of our model in advancing GUI visual agents.
****

Let's Chorus: Partner-aware Hybrid Song-Driven 3D Head Animation
Xiumei Xie, Zikai Huang, Wenhao Xu, Peng Xiao, Xuemiao Xu, Huaidong Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5467-5476
Singing is a vital form of human emotional expression and social interaction, distinguished from speech by its richer emotional nuances and freer expressive style. Thus, investigating 3D facial animation driven by singing holds significant research value. Our work focuses on 3D singing facial animation driven by mixed singing audio, and to the best of our knowledge, no prior studies have explored this area. Additionally, the absence of existing 3D singing datasets poses a considerable challenge. To address this, we collect a novel audiovisual dataset, ChorusHead which features synchronized mixed vocal audio and pseudo-3D flame motions for chorus singing. In addition, We propose a partner-aware 3D chorus head generation framework driven by mixed audio inputs. The proposed framework extracts emotional features from the background music and dependence between singers and models the head movement in a latent space from the Variational Autoencoder (VAE), enabling diverse interactive head animation generation. Extensive experimental results demonstrate that our approach effectively generates 3D facial animations of interacting singers, achieving notable improvements in realism and handling background music interference with strong robustness.
****

Twinner: Shining Light on Digital Twins in a Few Snaps
Jesus Zarzar, Tom Monnier, Roman Shapovalov, Andrea Vedaldi, David Novotny; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5859-5869
We present the first large reconstruction model, Twinner, capable of recovering a scene's illumination as well as an object's geometry and material properties from only a few posed images. Twinner is based on the Large Reconstruction Model and innovates in three key ways:1) We introduce a memory-efficient voxel-grid transformer whose memory scales only quadratically with the size of the voxel grid. 2) To deal with scarcity of high-quality ground-truth PBR-shaded models, we introduce a large fully-synthetic dataset of procedurally-generated PBR-textured objects lit with varied illumination.3) To narrow the synthetic-to-real gap, we finetune the model on real life datasets by means of a differentiable physically-based shading model, eschewing the need for ground-truth illumination or material properties which are challenging to obtain in real life. We demonstrate the efficacy of our model on the real life StanfordORB benchmark where, given few input views, we achieve reconstruction quality significantly superior to existing feed-forward reconstruction networks, and comparable to significantly slower per-scene optimization methods.
****

Infinity: Scaling Bitwise AutoRegressive Modeling for High-Resolution Image Synthesis
Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, Xiaobing Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15733-15744

We present Infinity, a Bitwise Visual AutoRegressive Modeling capable of generating high-resolution, photorealistic images following language instruction. Infinity refactors visual autoregressive model under a bitwise token prediction framework with an infinite-vocabulary classifier and bitwise self-correction mechanism. By theoretically expanding the tokenizer vocabulary size to infinity in Transformer, our method significantly unleashes powerful scaling capabilities to infinity compared to vanilla VAR. Extensive experiments indicate Infinity outperforms AutoRegressive Text-to-Image models by large margins, matches or surpasses leading diffusion models. Without extra optimization, Infinity generates a 1024x1024 image in 0.8s, 2.6xfaster than SD3-Medium, making it the fastest Text-to-Image model. Models and codes will be released to promote the further exploration of Infinity for visual generation.

****

DreamText: High Fidelity Scene Text Synthesis

Yibin Wang, Weizhong Zhang, Honghui Xu, Cheng Jin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28555-28563

Scene text synthesis involves rendering specified texts onto arbitrary images. Current methods typically formulate this task in an end-to-end manner but lack effective character-level guidance during training.Besides, their text encoders, pre-trained on a single font type, struggle to adapt to the diverse font styles encountered in practical applications.Consequently, these methods suffer from character distortion, repetition, and absence, particularly in polystylistic scenarios.To this end, this paper proposes DreamText for high-fidelity scene text synthesis.Our key idea is to reconstruct the diffusion training process, introducing more refined guidance tailored to this task, to expose and rectify the model's attention at the character level and strengthen its learning of text regions.This transformation poses a hybrid optimization challenge, involving both discrete and continuous variables. To effectively tackle this challenge, we employ a heuristic alternate optimization strategy. Meanwhile, we jointly train the text encoder and generator to comprehensively learn and utilize the diverse font present in the training dataset. This joint training is seamlessly integrated into the alternate optimization process, fostering a synergistic relationship between learning character embedding and re-estimating character attention.Specifically, in each step, we first encode potential character-generated position information from cross-attention maps into latent character masks. These masks are then utilized to update the representation of specific characters in the current step, which, in turn, enables the generator to correct the character's attention in the subsequent steps.Both qualitative and quantitative results demonstrate the superiority of our method to the state of the art.

****

MonoPlace3D: Learning 3D-Aware Object Placement for 3D Monocular Detection

Rishubh Parihar, Srinjay Sarkar, Sarthak Vora, Jogendra Nath Kundu, R. Venkatesh Babu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6531-6541

Current monocular 3D detectors are held back by the limited diversity and scale of real-world datasets. While data augmentation certainly helps, it's particularly difficult to generate realistic scene-aware augmented data for outdoor settings. Most current approaches to synthetic data generation focus on realistic object appearance through improved rendering techniques. However, we show that where and how objects are positioned is just as crucial for training effective 3D monocular detectors. The key obstacle lies in automatically determining realistic object placement parameters - including position, dimensions, and directional alignment when introducing synthetic objects into actual scenes. To address this, we introduce MonoPlace3D, a novel system that considers the 3D scene content to create realistic augmentations. Specifically, given a background scene, MonoPlace3D learns a distribution over plausible 3D bounding boxes. Subsequently, we render realistic objects and place them according to the locations sampled from the learned distribution. Our comprehensive evaluation on two standard datasets KITTI and NuScenes, demonstrates that MonoPlace3D significantly improves the accuracy of multiple existing monocular 3D detectors while being highly data efficient

.
****
HumanDreamer: Generating Controllable Human-Motion Videos via Decoupled Generation
Boyuan Wang, Xiaofeng Wang, Chaojun Ni, Guosheng Zhao, Zhiqin Yang, Zheng Zhu, Muyang Zhang, Yukun Zhou, Xinze Chen, Guan Huang, Lihong Liu, Xingang Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12391-12401
Human-motion video generation has been a challenging task, primarily due to the difficulty inherent in learning human body movements. While some approaches have attempted to drive human-centric video generation explicitly through pose control, these methods typically rely on poses derived from existing videos, thereby lacking flexibility. To address this, we propose HumanDreamer, a decoupled human video generation framework that first generates diverse poses from text prompts and then leverages these poses to generate human-motion videos. Specifically, we propose MotionVid, the largest dataset for human-motion pose generation. Based on the dataset, we present MotionDiT, which is trained to generate structured human-motion poses from text prompts. Besides, a novel LAMA loss is introduced, which together contribute to a significant improvement in FID by 62.4%, along with respective enhancements in R-precision for top1, top2, and top3 by 41.8%, 26.3%, and 18.3%, thereby advancing both the Text-to-Pose control accuracy and FID metrics. Our experiments across various Pose-to-Video baselines demonstrate that the poses generated by our method can produce diverse and high-quality human-motion videos. Furthermore, our model can facilitate other downstream tasks, such as pose sequence prediction and 2D-3D motion lifting.
****
ReVisionLLM: Recursive Vision-Language Model for Temporal Grounding in Hour-Long Videos
Tanveer Hannan, Md Mohaiminul Islam, Jindong Gu, Thomas Seidl, Gedas Bertasius; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19012-19022
Large language models (LLMs) excel at retrieving information from lengthy text, but their vision-language counterparts (VLMs) face difficulties with hour-long videos, especially for temporal grounding. Specifically, these VLMs are constrained by frame limitations, often losing essential temporal details needed for accurate event localization in extended video content. We propose ReVisionLLM, a recursive vision-language model designed to locate events in hour-long videos. Inspired by human search strategies, our model initially targets broad segments of interest, progressively revising its focus to pinpoint exact temporal boundaries. Our model can seamlessly handle videos of vastly different lengths--from minutes to hours. We also introduce a hierarchical training strategy that starts with short clips to capture distinct events and progressively extends to longer videos. To our knowledge, ReVisionLLM is the first VLM capable of temporal grounding in hour-long videos, outperforming previous state-of-the-art methods across multiple datasets by a significant margin (e.g., +2.6% R1@0.1 on MAD). The code is available in the supplementary and will be released.
****
ArtiFade: Learning to Generate High-quality Subject from Blemished Images
Shuya Yang, Shaozhe Hao, Yukang Cao, Kwan-Yee K. Wong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13167-13177
Subject-driven text-to-image generation has demonstrated remarkable advancements in its ability to learn and capture characteristics of a subject using only a limited number of images. However, existing methods commonly rely on high-quality images for training and often struggle to generate reasonable images when the input images are blemished by artifacts. This is primarily attributed to the inadequate capability of current techniques in distinguishing subject-related features from disruptive artifacts. In this paper, we introduce ArtiFade to tackle this issue and successfully generate high-quality artifact-free images from blemished datasets. Specifically, ArtiFade exploits fine-tuning of a pre-trained text-to-image model, aiming to remove artifacts. The elimination of artifacts is achie

ved by utilizing a specialized dataset that encompasses both unblemished images and their corresponding blemished counterparts during fine-tuning. ArtiFade also ensures the preservation of the original generative capabilities inherent within the diffusion model, thereby enhancing the overall performance of subject-driven methods in generating high-quality and artifact-free images. We further devise evaluation benchmarks tailored for this task. Through extensive qualitative and quantitative experiments, we demonstrate the generalizability of ArtiFade in effective artifact removal under both in-distribution and out-of-distribution scenarios.

****

SGCR: Spherical Gaussians for Efficient 3D Curve Reconstruction

Xinran Yang, Donghao Ji, Yuanqi Li, Jie Guo, Yanwen Guo, Junyuan Xie; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5793-5803

Neural rendering techniques have made substantial progress in generating photo-realistic 3D scenes. The latest 3D Gaussian Splatting technique has achieved high quality novel view synthesis as well as fast rendering speed. However, 3D Gaussians lack proficiency in defining accurate 3D geometric structures despite their explicit primitive representations. This is due to the fact that Gaussian's attributes are primarily tailored and fine-tuned for rendering diverse 2D images by their anisotropic nature. To pave the way for efficient 3D reconstruction, we present Spherical Gaussians, a simple and effective representation for 3D geometric boundaries, from which we can directly reconstruct 3D feature curves from a set of calibrated multi-view images. Spherical Gaussians is optimized from grid initialization with a view-based rendering loss, where a 2D edge map is rendered at a specific view and then compared to the ground-truth edge map extracted from the corresponding image, without the need for any 3D guidance or supervision. Given Spherical Gaussians serve as intermedia for the robust edge representation, we further introduce a novel optimization-based algorithm called SGCR to directly extract accurate parametric curves from aligned Spherical Gaussians. We demonstrate that SGCR outperforms existing state-of-the-art methods in 3D edge reconstruction while enjoying great efficiency.

****

Prompting Depth Anything for 4K Resolution Accurate Metric Depth Estimation

Haotong Lin, Sida Peng, Jingxiao Chen, Songyou Peng, Jiaming Sun, Minghuan Liu, Hujun Bao, Jiashi Feng, Xiaowei Zhou, Bingyi Kang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17070-17080

Prompts play a critical role in unleashing the power of language and vision foundation models for specific tasks. For the first time, we introduce prompting into depth foundation models, creating a new paradigm for metric depth estimation termed Prompt Depth Anything. Specifically, we use a low-cost LiDAR as the prompt to guide the Depth Anything model for accurate metric depth output, achieving up to 4K resolution. Our approach centers on a concise prompt fusion design that integrates the LiDAR at multiple scales within the depth decoder. To address training challenges posed by limited datasets containing both LiDAR depth and precise GT depth, we propose a scalable data pipeline that includes synthetic data LiDAR simulation and real data pseudo GT depth generation. Our approach sets new state-of-the-arts on the ARKitScenes and ScanNet++ datasets. Furthermore, we demonstrate that it benefits several downstream applications, including 3D reconstruction and generalized robotic grasping. Code will be released.

****

ProKeR: A Kernel Perspective on Few-Shot Adaptation of Large Vision-Language Models

Yassir Bendou, Amine Ouasfi, Vincent Gripon, Adnane Boukhayma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25092-25102

The growing popularity of Contrastive Language-Image Pretraining (CLIP) has led to its widespread application in various visual downstream tasks. To enhance CLIP's effectiveness and versatility, efficient few-shot adaptation techniques have been widely adopted. Among these approaches, training-free methods, particularl

y caching methods exemplified by Tip-Adapter, have gained attention for their li ghtweight adaptation without the need for additional fine-tuning. In this paper, we revisit Tip-Adapter from a kernel perspective, showing that caching methods function as local adapters and are connected to a well-established kernel litera ture. Drawing on this insight, we offer a theoretical understanding of how these methods operate and suggest multiple avenues for enhancing the Tip-Adapter base line. Notably, our analysis shows the importance of incorporating global informa tion in local adapters. Therefore, we subsequently propose a global method that learns a proximal regularizer in a reproducing kernel Hilbert space (RKHS) using CLIP as a base learner. Our method, which we call ProKeR (Proximal Kernel ridge Regression), has a closed form solution and achieves state-of-the-art performan ces across 11 datasets in the standard few-shot adaptation benchmark. Code is av ailable at https://ybendou.github.io/ProKeR/.

****

Advancing Semantic Future Prediction through Multimodal Visual Sequence Transfor mers

Efstathios Karypidis, Ioannis Kakogeorgiou, Spyros Gidaris, Nikos Komodakis; Pro ceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3793-3803

Semantic future prediction is important for autonomous systems navigating dynami c environments. This paper introduces FUTURIST, a method for multimodal future s emantic prediction that uses a unified and efficient visual sequence transformer architecture. Our approach incorporates a multimodal masked visual modeling obj ective and a novel masking mechanism designed for multimodal training. This allo ws the model to effectively integrate visible information from various modalitie s, improving prediction accuracy. Additionally, we propose a VAE-free hierarchic al tokenization process, which reduces computational complexity, streamlines the training pipeline, and enables end-to-end training with high-resolution, multim odal inputs. We validate FUTURIST on the Cityscapes dataset, demonstrating state -of-the-art performance in future semantic segmentation for both short- and mid- term forecasting. We provide the implementation code and model weights at https: //github.com/Sta8is/FUTURIST.

****

GET: Unlocking the Multi-modal Potential of CLIP for Generalized Category Discov ery

Enguang Wang, Zhimao Peng, Zhengyuan Xie, Fei Yang, Xialei Liu, Ming-Ming Cheng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2 025, pp. 20296-20306

Given unlabelled datasets containing both old and new categories, generalized ca tegory discovery (GCD) aims to accurately discover new classes while correctly c lassifying old classes. Current GCD methods only use a single visual modality of information, resulting in poor classification of visually similar classes. As a different modality, text information can provide complementary discriminative i nformation, which motivates us to introduce it into the GCD task. However, the l ack of class names for unlabelled data makes it impractical to utilize text info rmation. To tackle this challenging problem, in this paper, we propose a Text Em bedding Synthesizer (TES) to generate pseudo text embeddings for unlabelled samp les. Specifically, our TES leverages the property that CLIP can generate aligned vision-language features, converting visual embeddings into tokens of the CLIP' s text encoder to generate pseudo text embeddings. Besides, we employ a dual-bra nch framework, through the joint learning and instance consistency of different modality branches, visual and semantic information mutually enhance each other, promoting the interaction and fusion of visual and text knowledge. Our method un locks the multi-modal potentials of CLIP and outperforms the baseline methods by a large margin on all GCD benchmarks, achieving new state-of-the-art.

****

On the Out-Of-Distribution Generalization of Large Multimodal Models

Xingxuan Zhang, Jiansheng Li, Wenjing Chu, junjia hai, Renzhe Xu, Yuqing Yang, S hikai Guan, Jiazheng Xu, Liping Jing, Peng Cui; Proceedings of the Computer Visi on and Pattern Recognition Conference (CVPR), 2025, pp. 10315-10326

We investigate the generalization boundaries of current Large Multimodal Models (LMMs) via comprehensive evaluation under out-of-distribution scenarios and doma in-specific tasks. We evaluate their zero-shot generalization across synthetic i mages, real-world distributional shifts, and specialized datasets like medical a nd molecular imagery. Empirical results indicate that LMMs struggle with general ization beyond common training domains, limiting their direct application withou t adaptation. To understand the cause of unreliable performance, we analyze thre e hypotheses: semantic misinterpretation, visual feature extraction insufficienc y, and mapping deficiency. Results identify mapping deficiency as the primary hu rdle. To address this problem, we show that in-context learning (ICL) can signif icantly enhance LMMs' generalization, opening new avenues for overcoming general ization barriers. We further explore the robustness of ICL under distribution sh ifts and show its vulnerability to domain shifts, label shifts, and spurious cor relation shifts between in-context examples and test data.

****

Enhanced Contrastive Learning with Multi-view Longitudinal Data for Chest X-ray Report Generation

Kang Liu, Zhuoqi Ma, Xiaolu Kang, Yunan Li, Kun Xie, Zhicheng Jiao, Qiguang Miao ; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10348-10359

Automated radiology report generation offers an effective solution to alleviate radiologists' workload. However, most existing methods focus primarily on single or fixed-view images to model current disease conditions, which limits diagnost ic accuracy and overlooks disease progression. Although some approaches utilize longitudinal data to track disease progression, they still rely on single images to analyze current visits. To address these issues, we propose enhanced contras tive learning with Multi-view Longitudinal data to facilitate chest X-ray Report Generation, named MLRG. Specifically, we introduce a multi-view longitudinal co ntrastive learning method that integrates spatial information from current multi -view images and temporal information from longitudinal data. This method also u tilizes the inherent spatiotemporal information of radiology reports to supervis e the pre-training of visual and textual representations. Subsequently, we prese nt a tokenized absence encoding technique to flexibly handle missing patient-spe cific prior knowledge, allowing the model to produce more accurate radiology rep orts based on available prior knowledge. Extensive experiments on MIMIC-CXR, MIM IC-ABN, and Two-view CXR datasets demonstrate that our MLRG outperforms recent s tate-of-the-art methods, achieving a 2.3% BLEU-4 improvement on MIMIC-CXR, a 5.5 % F1 score improvement on MIMIC-ABN, and a 2.7% F1 RadGraph improvement on Two-v iew CXR.

****

MonoTAKD: Teaching Assistant Knowledge Distillation for Monocular 3D Object Dete ction

Hou-I Liu, Christine Wu, Jen-Hao Cheng, Wenhao Chai, Shian-Yun Wang, Gaowen Liu, Hugo Latapie, Jhih-Ciang Wu, Jenq-Neng Hwang, Hong-Han Shuai, Wen-Huang Cheng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 20 25, pp. 22266-22275

Monocular 3D object detection (Mono3D) holds noteworthy promise for autonomous d riving applications owing to the cost-effectiveness and rich visual context of m onocular camera sensors. However, depth ambiguity poses a significant challenge, as it requires extracting precise 3D scene geometry from a single image, result ing in suboptimal performance when transferring knowledge from a LiDAR-based tea cher model to a camera-based student model. To facilitate effective distillation , we introduce Monocular Teaching Assistant Knowledge Distillation (MonoTAKD), w hich proposes a camera-based teaching assistant (TA) model to transfer robust 3D visual knowledge to the student model, leveraging the smaller feature represent ation gap. Additionally, we define 3D spatial cues as residual features that cap ture the differences between the teacher and the TA models. We then leverage the se cues to improve the student model's 3D perception capabilities. Experimental results show that our MonoTAKD achieves state-of-the-art performance on the KITT I3D dataset. Furthermore, we evaluate the performance on nuScenes and KITTI raw

datasets to demonstrate the generalization of our model to multi-view 3D and uns
upervised data settings. Our code is available at https://github.com/hoiliu-0801
/MonoTAKD.
****

Test-Time Domain Generalization via Universe Learning: A Multi-Graph Matching Ap
proach for Medical Image Segmentation
Xingguo Lv, Xingbo Dong, Liwen Wang, Jiewen Yang, Lei Zhao, Bin Pu, Zhe Jin, Xue
jun Li; Proceedings of the Computer Vision and Pattern Recognition Conference (C
VPR), 2025, pp. 15621-15631
Despite domain generalization (DG) has significantly addressed the performance d
egradation of pre-trained models caused by domain shifts, it often falls short i
n real-world deployment. Test-time adaptation (TTA), which adjusts a learned mod
el using unlabeled test data, presents a promising solution. However, most exist
ing TTA methods struggle to deliver strong performance in medical image segmenta
tion, primarily because they overlook the crucial prior knowledge inherent to me
dical images. To address this challenge, we incorporate morphological informatio
n and propose a framework based on multi-graph matching. Specifically, we introd
uce learnable universe embeddings that integrate morphological priors during mul
ti-source training, along with novel unsupervised test-time paradigms for domain
 adaptation. This approach guarantees cycle-consistency in multi-matching while
enabling the model to more effectively capture the invariant priors of unseen da
ta, significantly mitigating the effects of domain shifts. Extensive experiments
 demonstrate that our method outperforms other state-of-the-art approaches on tw
o medical image segmentation benchmarks for both multi-source and single-source
domain generalization tasks. The source code is available at https://github.com/
Yore0/TTDG-MGM.
****

Easy-editable Image Vectorization with Multi-layer Multi-scale Distributed Visua
l Feature Embedding
Ye Chen, Zhangli Hu, Zhongyin Zhao, Yupeng Zhu, Yue Shi, Yuxuan Xiong, Bingbing
Ni; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)
, 2025, pp. 23345-23354
Current parameterized image representations embed visual information along the s
emantic boundaries and struggle to express the internal detailed texture structu
res of image components, leading to a lack of content consistency after image ed
iting and driving. To address these challenges, this work proposes a novel param
eterized representation based on hierarchical image proxy geometry, utilizing m
ulti-layer hierarchically interrelated proxy geometric control points to embed m
ulti-scale long-range structures and fine-grained texture details. The proposed
representation enables smoother and more continuous interpolation during image r
endering and ensures high-quality consistency within image components during ima
ge editing. Additionally, under the layer-wise representation strategy based on
semantic-aware image layer decomposition, we enable decoupled image shape/textur
e editing of the targets of interest within the image. Extensive experimental re
sults on image vectorization and editing tasks demonstrate that our proposed met
hod achieves high rendering accuracy of general images, including natural images
, with a significantly higher image parameter compression ratio, facilitating us
er-friendly editing of image semantic components.
****

Acquire and then Adapt: Squeezing out Text-to-Image Model for Image Restoration
Junyuan Deng, Xinyi Wu, Yongxing Yang, Congchao Zhu, Song Wang, Zhenyao Wu; Proc
eedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025,
pp. 23195-23206
Recently, pre-trained text-to-image (T2I) models have been extensively adopted f
or real-world image restoration because of their powerful generative prior. Howe
ver, controlling these large models for image restoration usually requires a lar
ge number of high-quality images and immense computational resources for trainin
g, which is costly and not privacy-friendly. In this paper, we find that the wel
l-trained large T2I model (i.e., Flux) is able to produce a variety of high-qual
ity images aligned with real-world distributions, offering an unlimited supply o

f training samples to mitigate the above issue. Specifically, we proposed a trai
ning data construction pipeline for image restoration, namely FluxGen, which inc
ludes unconditional image generation, image selection, and degraded image simula
tion. A novel light-weighted adapter (FluxIR) with squeeze-and-excitation layer
s is also carefully designed to control the large Diffusion Transformer (DiT)-ba
sed T2I model so that reasonable details can be restored. Experiments demonstrat
e that our proposed method enables the Flux model to adapt effectively to real-w
orld image restoration tasks, achieving superior scores and visual quality on bo
th synthetic and real-world degradation datasets - at only about 8.5% of the tra
ining cost compared to current approaches.
****

DeDe: Detecting Backdoor Samples for SSL Encoders via Decoders
Sizai Hou, Songze Li, Duanyi Yao; Proceedings of the Computer Vision and Pattern
 Recognition Conference (CVPR), 2025, pp. 20675-20684
Self-supervised learning (SSL) is pervasively exploited in training high-quality
 upstream encoders with a large amount of unlabeled data. However, it is found t
o be susceptible to backdoor attacks merely via polluting a small portion of tra
ining data. The victim encoders associate triggered inputs with target embedding
s, e.g., mapping a triggered cat image to an airplane embedding, such that the d
ownstream tasks inherit unintended behaviors when the trigger is activated. Emer
ging backdoor attacks have shown great threats across different SSL paradigms su
ch as contrastive learning and CLIP, yet limited research is devoted to defendin
g against such attacks, and existing defenses fall short in detecting advanced s
tealthy backdoors. To address the limitations, we propose a novel detection mech
anism, DeDe, which detects the activation of backdoor mappings caused by trigger
ed inputs on victim encoders. Specifically, DeDe trains a decoder for any given
SSL encoder using an auxiliary dataset (which can be out-of-distribution or even
 slightly poisoned), so that for any triggered input that misleads the encoder i
nto the target embedding, the decoder generates an output image significantly di
fferent from the input. DeDe leverages the discrepancy between the input and the
 decoded output to identify potential backdoor misbehavior during inference. We
empirically evaluate DeDe on both contrastive learning and CLIP models against v
arious types of backdoor attacks. Our results demonstrate promising detection ef
fectiveness over various advanced attacks and superior performance compared over
 state-of-the-art detection methods.
****

Towards Scalable Human-aligned Benchmark for Text-guided Image Editing
Suho Ryu, Kihyun Kim, Eugene Baek, Dongsoo Shin, Joonseok Lee; Proceedings of th
e Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18292-183
01
A variety of text-guided image editing models have been proposed recently. Howev
er, there is no widely-accepted standard evaluation method mainly due to the sub
jective nature of the task, letting researchers rely on manual user study. To ad
dress this, we introduce a novel Human-Aligned benchmark for Text-guided Image E
diting (HATIE). Providing a large-scale benchmark set covering a wide range of e
diting tasks, it allows reliable evaluation, not limited to specific easy-to-eva
luate cases. Also, HATIE provides a fully-automated and omnidirectional evaluati
on pipeline. Particularly, we combine multiple scores measuring various aspects
of editing so as to align with human perception. We empirically verify that the
evaluation of HATIE is indeed human-aligned in various aspects, and provide benc
hmark results on several state-of-the-art models to provide deeper insights on t
heir performance.
****

Devils in Middle Layers of Large Vision-Language Models: Interpreting, Detecting
 and Mitigating Object Hallucinations via Attention Lens
Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, Xu Yang; Procee
dings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp
. 25004-25014
Hallucinations in Large Vision-Language Models (LVLMs) significantly undermine t
heir reliability, motivating researchers to explore the causes of hallucination.

However, most studies primarily focus on the language aspect rather than the visual. In this paper, we address how LVLMs process visual information and whether this process causes hallucination. Firstly, we use the attention lens to identify the stages at which LVLMs handle visual data, discovering that the middle layers are crucial. Moreover, we find that these layers can be further divided into two stages: "visual information enrichment" and "semantic refinement" which respectively propagate visual data to object tokens and interpret it through text. By analyzing attention patterns during the visual information enrichment stage, we find that real tokens consistently receive higher attention weights than hallucinated ones, serving as a strong indicator of hallucination. Further examination of multi-head attention maps reveals that hallucination tokens often result from heads interacting with inconsistent objects. Based on these insights, we propose a simple inference-time method that adjusts visual attention by integrating information across various heads. Extensive experiments demonstrate that this approach effectively mitigates hallucinations in mainstream LVLMs without additional training costs.
****

SpectroMotion: Dynamic 3D Reconstruction of Specular Scenes
Cheng-De Fan, Chen-Wei Chang, Yi-Ruei Liu, Jie-Ying Lee, Jiun-Long Huang, Yu-Chee Tseng, Yu-Lun Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21328-21338
We present SpectroMotion, a novel approach that combines 3D Gaussian Splatting (3DGS) with physically-based rendering (PBR) and deformation fields to reconstruct dynamic specular scenes. Previous methods extending 3DGS to model dynamic scenes have struggled to represent specular surfaces accurately. Our method addresses this limitation by introducing a residual correction technique for accurate surface normal computation during deformation, complemented by a deformable environment map that adapts to time-varying lighting conditions. We implement a coarse-to-fine training strategy significantly enhancing scene geometry and specular color prediction. It is the only existing 3DGS method capable of synthesizing photorealistic real-world dynamic specular scenes, outperforming state-of-the-art methods in rendering complex, dynamic, and specular scenes. Please see our project page at https://cdfan0627.github.io/spectromotion/.
****

Scaling Inference Time Compute for Diffusion Models
Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, Saining Xie; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2523-2534
Generative models have made significant impacts across various domains, largely due to their ability to scale during training by increasing data, computational resources, and model size, a phenomenon characterized by the scaling laws. Recent research has begun to explore inference-time scaling behavior in Large Language Models (LLMs), revealing how performance can further improve with additional computation during inference. Unlike LLMs, diffusion models inherently possess the flexibility to adjust inference-time computation via the number of denoising steps, although the performance gains typically flatten after a few dozen. In this work, we explore the inference-time scaling behavior of diffusion models beyond increasing denoising steps and investigate how the generation performance can further improve with increased computation. Specifically, we consider a search problem aimed at identifying better noises for the diffusion sampling process. We structure the design space along two axes: the verifiers used to provide feedback, and the algorithms used to find better noise candidates. Through extensive experiments on class-conditioned and text-conditioned image generation benchmarks, our findings reveal that increasing inference-time compute leads to substantial improvements in the quality of samples generated by diffusion models, and with the complicated nature of images, combinations of the components in the framework can be specifically chosen to conform with different application scenario.
****

Coeff-Tuning: A Graph Filter Subspace View for Tuning Attention-Based Large Models

Zichen Miao, Wei Chen, Qiang Qiu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20146-20157

Transformer-based large pre-trained models have shown remarkable generalization ability, and various parameter-efficient fine-tuning (PEFT) methods have been proposed to customize these models on downstream tasks with minimal computational and memory budgets. Previous PEFT methods are primarily designed from a tensor-decomposition perspective that tries to effectively tune the linear transformation by finding the smallest subset of parameters to train. Our study adopts an orthogonal view by representing the attention operation as a graph convolution and formulating the multi-head attention maps as a convolutional filter subspace, with each attention map as a subspace element. In this paper, we propose to tune the large pre-trained transformers by learning a small set of combination coefficients that construct a more expressive filter subspace from the original multi-head attention maps. We show analytically and experimentally that the tuned filter subspace can effectively expand the feature space of the multi-head attention and further enhance the capacity of transformers. We further stabilize the fine-tuning with a residual parameterization of the tunable subspace coefficients, and enhance the generalization with a regularization design by directly applying dropout on the tunable coefficient during training. The tunable coefficients take a tiny number of parameters and can be combined with previous PEFT methods in a plug-and-play manner. Extensive experiments show that our approach achieves superior performances than PEFT baselines with neglectable additional parameters.

****

VTON 360: High-Fidelity Virtual Try-On from Any Viewing Direction

Zijian He, Yuwei Ning, Yipeng Qin, Guangrun Wang, Sibei Yang, Liang Lin, Guanbin Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26388-26398

Virtual Try-On (VTON) is a transformative technology in e-commerce and fashion design, enabling realistic digital visualization of clothing on individuals. In this work, we propose VTON 360, a novel 3D VTON method that addresses the open challenge of achieving high-fidelity VTON that supports any-view rendering. Specifically, we leverage the equivalence between a 3D model and its rendered multi-view 2D images, and reformulate 3D VTON as an extension of 2D VTON that ensures 3D consistent results across multiple views.To achieve this, we extend 2D VTON models to include multi-view garments and clothing-agnostic human body images as input, and propose several novel techniques to enhance them, including: i) a pseudo-3D pose representation using normal maps derived from the SMPL-X 3D human model, ii) a multi-view spatial attention mechanism that models the correlations between features from different viewing angles, and iii) a multi-view CLIP embedding that enhances the garment CLIP features used in 2D VTON with camera information. Extensive experiments on large-scale real datasets and clothing images from e-commerce platforms demonstrate the effectiveness of our approach.

****

MVBoost: Boost 3D Reconstruction with Multi-View Refinement

Xiangyu Liu, Xiaomei Zhang, Zhiyuan Ma, Xiangyu Zhu, Zhen Lei; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21664-21673

Recent advancements in 3D object reconstruction have been remarkable, yet most current 3D models rely heavily on existing 3D datasets. The scarcity of diverse 3D datasets results in limited generalization capabilities of 3D reconstruction models. In this paper, we propose a novel framework for boosting 3D reconstruction with multi-view refinement (MVBoost) by generating pseudo-GT data. The key of MVBoost is combining the advantages of the high accuracy of the multi-view generation model and the consistency of the 3D reconstruction model to create a reliable data source. Specifically, given a single-view input image, we employ a multi-view diffusion model to generate multiple views, followed by a large 3D reconstruction model to produce consistent 3D data. MVBoost then adaptively refines these multi-view images, rendered from the consistent 3D data, to build a large-scale multi-view dataset for training a feed-forward 3D reconstruction model. Additionally, the input view optimization is designed to optimize the corresponding

viewpoints based on the user's input image, ensuring that the most important vie
wpoint is accurately tailored to the user's needs. Extensive evaluations demonst
rate that our method achieves superior reconstruction results and robust general
ization compared to prior works.
****

Chat-based Person Retrieval via Dialogue-Refined Cross-Modal Alignment
Yang Bai, Yucheng Ji, Min Cao, Jinqiao Wang, Mang Ye; Proceedings of the Compute
r Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3952-3962
Traditional text-based person retrieval (TPR) relies on a single-shot text as qu
ery to retrieve the target person, assuming that the query completely captures t
he user's search intent. However, in real-world scenarios, it can be challenging
 to ensure the information completeness of such single-shot text. To address thi
s limitation, we propose chat-based person retrieval (ChatPR), a new paradigm th
at takes an interactive dialogue as query to perform the person retrieval, engag
ing the user in conversational context to progressively refine the query for acc
urate person retrieval. The primary challenge in ChatPR is the lack of available
 dialogue-image paired data. To overcome this challenge, we establish ChatPedes,
 the first dataset designed for ChatPR, which is constructed by leveraging large
 language models to automate the question generation and simulate user responses
. Additionally, to bridge the modality gap between dialogues and images, we prop
ose a dialogue-refined cross-modal alignment (DiaNA) framework, which leverages
two adaptive attribute refiners to bottleneck the conversational and visual info
rmation for fine-grained cross-modal alignment. Moreover, we propose a dialogue-
specific data augmentation strategy, random round retaining, to further enhance
the model's generalization ability across varying dialogue lengths. Extensive ex
periments demonstrate that DiaNA significantly outperforms existing TPR approach
es, highlighting the effectiveness of conversational interactions for person ret
rieval.
****

Category-Agnostic Neural Object Rigging
Guangzhao He, Chen Geng, Shangzhe Wu, Jiajun Wu; Proceedings of the Computer Vis
ion and Pattern Recognition Conference (CVPR), 2025, pp. 22078-22088
The motion of deformable 4D objects lies in a low-dimensional manifold. To bette
r capture the low dimensionality and enable better controllability, traditional
methods have devised several heuristic-based methods, i.e., rigging, for manipul
ating dynamic objects in an intuitive fashion. However, such representations are
 not scalable due to the need for expert knowledge of specific categories. Inste
ad, we study the automatic exploration of such low-dimensional structures in a p
urely data-driven manner. Specifically, we design a novel representation that en
codes deformable 4D objects into a sparse set of spatially grounded blobs and an
 instance-aware feature volume to disentangle the pose and instance information
of the 3D shape. With such a representation, we can manipulate the pose of 3D ob
jects intuitively by modifying the parameters of the blobs, while preserving ric
h instance-specific information. We evaluate the proposed method on a variety of
 object categories and demonstrate the effectiveness of the proposed framework.
Project page: https://guangzhaohe.com/canor.
****

AVF-MAE++: Scaling Affective Video Facial Masked Autoencoders via Efficient Audi
o-Visual Self-Supervised Learning
Xuecheng Wu, Heli Sun, Yifan Wang, Jiayu Nie, Jie Zhang, Yabing Wang, Junxiao Xu
e, Liang He; Proceedings of the Computer Vision and Pattern Recognition Conferen
ce (CVPR), 2025, pp. 9142-9153
Affective Video Facial Analysis (AVFA) is important for advancing emotion-aware
AI, yet the persistent data scarcity in AVFA presents challenges. Recently, the
self-supervised learning (SSL) technique of Masked Autoencoders (MAE) has gained
 significant attention, particularly in its audio-visual adaptation. Insights fr
om general domains suggest that scaling is vital for unlocking impressive improv
ements, though its effects on AVFA remain largely unexplored. Additionally, capt
uring both intra- and inter-modal correlations through scalable representations
is a crucial challenge in this field. To tackle these gaps, we introduce AVF-MAE

++, a series audio-visual MAE designed to explore the impact of scaling on AVFA with a focus on advanced correlation modeling. Our method incorporates a novel audio-visual dual masking strategy and an improved modality encoder with a holistic view to better support scalable pre-training. Furthermore, we propose the Iteratively Audio-Visual Correlations Learning Module to improve correlations capture within the SSL framework, bridging the limitations of prior methods. To support smooth adaptation and mitigate overfitting, we also introduce a progressive semantics injection strategy, which structures training in three stages. Extensive experiments across 17 datasets, spanning three key AVFA tasks, demonstrate the superior performance of AVF-MAE++, establishing new state-of-the-art outcomes. Ablation studies provide further insights into the critical design choices driving these gains. Code is released at https://github.com/XuecWu/AVF-MAE.

****

POPEN: Preference-Based Optimization and Ensemble for LVLM-Based Reasoning Segmentation

Lanyun Zhu, Tianrun Chen, Qianxiong Xu, Xuanyi Liu, Deyi Ji, Haiyang Wu, De Wen Soh, Jun Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30231-30240

Existing LVLM-based reasoning segmentation methods often suffer from imprecise segmentation results and hallucinations in their text responses. This paper introduces POPEN, a novel framework designed to address these issues and achieve improved results. POPEN includes a preference-based optimization method to finetune the LVLM, aligning it more closely with human preferences and thereby generating better text responses and segmentation results. Additionally, POPEN introduces a preference-based ensemble method for inference, which integrates multiple outputs from the LVLM using a preference-score-based attention mechanism for refinement. To better adapt to the segmentation task, we incorporate several task-specific designs in our POPEN framework, including a new approach for collecting segmentation preference data with a curriculum learning mechanism, and a novel preference optimization loss to refine the segmentation capability of the LVLM. Experiments demonstrate that our method achieves state-of-the-art performance in reasoning segmentation, exhibiting minimal hallucination in text responses and the highest segmentation accuracy compared to previous advanced methods like LISA and PixelLM.

****

DiffusionSfM: Predicting Structure and Motion via Ray Origin and Endpoint Diffusion

Qitao Zhao, Amy Lin, Jeff Tan, Jason Y. Zhang, Deva Ramanan, Shubham Tulsiani; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6317-6326

Current Structure-from-Motion (SfM) methods typically follow a two-stage pipeline, combining learned or geometric pairwise reasoning with a subsequent global optimization step. In contrast, we propose a data-driven multi-view reasoning approach that directly infers 3D scene geometry and camera poses from multi-view images. Our framework, DiffusionSfM, parameterizes scene geometry and cameras as pixel-wise ray origins and endpoints in a global frame and employs a transformer-based denoising diffusion model to predict them from multi-view inputs. To address practical challenges in training diffusion models with missing data and unbounded scene coordinates, we introduce specialized mechanisms that ensure robust learning. We empirically validate DiffusionSfM on both synthetic and real datasets, demonstrating that it outperforms classical and learning-based approaches while naturally modeling uncertainty.

****

GroundingFace: Fine-grained Face Understanding via Pixel Grounding Multimodal Large Language Model

Yue Han, Jiangning Zhang, Junwei Zhu, Runze Hou, Xiaozhong Ji, Chuming Lin, Xiaobin Hu, Zhucun Xue, Yong Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3942-3951

Multimodal Language Learning Models (MLLMs) have shown remarkable performance in image understanding, generation, and editing, with recent advancements achievin

g pixel-level grounding with reasoning. However, these models for common objects struggle with fine-grained face understanding. In this work, we introduce the FacePlayGround-240K dataset, the first pioneering large-scale, pixel-grounded face caption and question-answer (QA) dataset, meticulously curated for alignment pretraining and instruction-tuning. We present the GroundingFace framework, specifically designed to enhance fine-grained face understanding. This framework significantly augments the capabilities of existing grounding models in face part segmentation, face attribute comprehension, while preserving general scene understanding. Comprehensive experiments validate that our approach surpasses current state-of-the-art models in pixel-grounded face captioning/QA and various downstream tasks, including face captioning, referring segmentation, and zero-shot face attribute recognition.

****

Self-Supervised Cross-View Correspondence with Predictive Cycle Consistency

Alan Baade, Changan Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16753-16763

Learning self-supervised visual correspondence is a long-studied task fundamental to visual understanding and human perception. However, existing correspondence methods largely focus on small image transformations, such as object tracking in high-framerate videos or learning pixel-to-pixel mappings between images with high view overlap. This severely limits their application in dynamic multi-view settings such as robot imitation learning. In this work, we introduce Predictive Cycle Consistency for learning object correspondence between extremely disjoint views of a scene without paired segmentation data. Our technique bootstraps object correspondence pseudolabels from raw image segmentations using conditional grayscale colorization and a cycle-consistency refinement prior. We then train deep ViTs on these pseudolabels, which we use to generate higher-quality pseudolabels and iteratively train better correspondence models. We demonstrate the performance of our method under both extreme in-the-wild camera view changes and across large temporal gaps in video. Our approach beats all prior supervised and prior SoTA self-supervised correspondence models on the EgoExo4D correspondence benchmark (+6.7 IoU Exo Query) and the prior SoTA self-supervised methods SiamMAE and DINO V1&V2 on the DAVIS-2017 and LVOS datasets across large frame gaps.

****

MMAudio: Taming Multimodal Joint Training for High-Quality Video-to-Audio Synthesis

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, Yuki Mitsufuji; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28901-28911

We propose to synthesize high-quality and synchronized audio, given video and optional text conditions, using a novel multimodal joint training framework (MMAudio). In contrast to single-modality training conditioned on (limited) video data only, MMAudio is jointly trained with larger-scale, readily available text-audio data to learn to generate semantically aligned high-quality audio samples. Additionally, we improve audio-visual synchrony with a conditional synchronization module that aligns video conditions with audio latents at the frame level. Trained with a flow matching objective, MMAudio achieves new video-to-audio state-of-the-art among public models in terms of audio quality, semantic alignment, and audio-visual synchronization, while having a low inference time (1.23s to generate an 8s clip) and just 157M parameters. MMAudio also achieves surprisingly competitive performance in text-to-audio generation, showing that joint training does not hinder single-modality performance. Code, models, and demo are available at: hkchengrex.github.io/MMAudio

****

CryptoFace: End-to-End Encrypted Face Recognition

Wei Ao, Vishnu Naresh Boddeti; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19197-19206

Face recognition is central to many authentication, security, and personalized applications. Yet, it suffers from significant privacy risks, particularly arising from unauthorized access to sensitive biometric data. This paper introduces Cr

yptoFace, the first end-to-end encrypted face recognition system with fully homo morphic encryption (FHE). It enables secure processing of facial data across all stages of a face-recognition process--feature extraction, storage, and matching --without exposing raw images or features. We introduce a mixture of shallow pat ch convolutional networks to support higher-dimensional tensors via patch-based processing while reducing the multiplicative depth and thus inference latency. P arallel FHE evaluation of these networks ensures near-resolution-independent lat ency. On standard face recognition benchmarks, CryptoFace significantly accelera tes inference and increases verification accuracy compared to the state-of-the-a rt FHE neural networks adapted for face recognition. CryptoFace will facilitate secure face recognition systems requiring robust and provable security. The code is available at https://github.com/human-analysis/CryptoFace.

****

Relation-Rich Visual Document Generator for Visual Information Extraction

Zi-Han Jiang, Chien-Wei Lin, Wei-Hua Li, Hsuan-Tung Liu, Yi-Ren Yeh, Chu-Song Ch en; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) , 2025, pp. 14449-14459

Despite advances in Large Language Models (LLMs) and Multimodal LLMs (MLLMs) for visual document understanding (VDU), visual information extraction (VIE) from r elation-rich documents remains challenging due to the layout diversity and limit ed training data. While existing synthetic document generators attempt to addres s data scarcity, they either rely on manually designed layouts and templates, or adopt rule-based approaches that limit layout diversity. Besides, current layou t generation methods focus solely on topological patterns without considering te xtual content, making them impractical for generating documents with complex ass ociations between the contents and layouts. In this paper, we propose a Relation -rIch visual Document GEnerator (RIDGE) that addresses these limitations through a two-stage approach: (1) Content Generation, which leverages LLMs to generate document content using a carefully designed Hierarchical Structure Text format w hich captures entity categories and relationships, and (2) Content-driven Layout Generation, which learns to create diverse, plausible document layouts solely f rom easily available Optical Character Recognition (OCR) results, requiring no h uman labeling or annotations efforts. Experimental results have demonstrated tha t our method significantly enhances the performance of document understanding mo dels on various VIE benchmarks.

****

DynFocus: Dynamic Cooperative Network Empowers LLMs with Video Understanding

Yudong Han, Qingpei Guo, Liyuan Pan, Liu Liu, Yu Guan, Ming Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8512-8 522

The challenge in LLM-based video understanding lies in preserving visual and sem antic information in long videos while maintaining a memory-affordable token cou nt. However, redundancy and correspondence in videos have hindered the performan ce potential of existing methods. Through statistical learning on current datase ts, we observe that redundancy occurs in both repeated and answer-irrelevant fra mes, and the corresponding frames vary with different questions. This suggests t he possibility of adopting dynamic encoding to balance detailed video informatio n preservation with token budget reduction. To this end, we propose a dynamic co operative network, DynFocus, for memory-efficient video encoding in this paper. Specifically, i) a Dynamic Event Prototype Estimation (DPE) module to dynamicall y select meaningful frames for question answering; (ii) a Compact Cooperative En coding (CCE) module that encodes meaningful frames with detailed visual appearan ce and the remaining frames with sketchy perception separately. We evaluate our method on five publicly available benchmarks, and experimental results consisten tly demonstrate that our method achieves competitive performance.

****

Mimic In-Context Learning for Multimodal Tasks

Yuchu Jiang, Jiale Fu, Chenduo Hao, Xinting Hu, Yingzhe Peng, Xin Geng, Xu Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2 025, pp. 29825-29835

Recently, In-context Learning (ICL) has become a significant inference paradigm in Large Multimodal Models (LMMs), utilizing a few in-context demonstrations (ICDs) to prompt LMMs for new tasks. However, the synergistic effects in multimodal data increase the sensitivity of ICL performance to the configurations of ICDs, stimulating the need for a more stable and general mapping function. Mathematically, in Transformer-based models, ICDs act as "shift vectors" added to the hidden states of query tokens. Inspired by this, we introduce Mimic In-Context Learning (MimIC) to learn stable and generalizable shift effects from ICDs. Specifically, compared with some previous shift vector-based methods, MimIC more strictly approximates the shift effects by integrating lightweight learnable modules into LMMs with four key enhancements: 1) inserting shift vectors after attention layers, 2) assigning a shift vector to each attention head, 3) making shift magnitude query-dependent, and 4) employing a layer-wise alignment loss. Extensive experiments on two LMMs (Idefics-9b and Idefics2-8b-base) across three multimodal tasks (VQAv2, OK-VQA, Captioning) demonstrate that MimIC outperforms existing shift vector-based methods. The code is available at https://anonymous.4open.science/r/MimIC/.

****

PromptHash:Affinity-Prompted Collaborative Cross-Modal Learning for Adaptive Hashing Retrieval

Qiang Zou, Shuli Cheng, Jiayi Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19649-19658

Cross-modal hashing is a promising approach for efficient data retrieval and storage optimization. However, contemporary methods exhibit significant limitations in semantic preservation, contextual integrity, and information redundancy, which constrains retrieval efficacy. We present PromptHash, an innovative framework leveraging affinity prompt-aware collaborative learning for adaptive cross-modal hashing. We propose an end-to-end framework for affinity-prompted collaborative hashing, with the following fundamental technical contributions: (i) a text affinity prompt learning mechanism that preserves contextual information while maintaining parameter efficiency, (ii) an adaptive gated selection fusion architecture that synthesizes State Space Model with Transformer network for precise cross-modal feature integration, and (iii) a prompt affinity alignment strategy that bridges modal heterogeneity through hierarchical contrastive learning. To the best of our knowledge, this study presents the first investigation into affinity prompt awareness within collaborative cross-modal adaptive hash learning, establishing a paradigm for enhanced semantic consistency across modalities. Through comprehensive evaluation on three benchmark multi-label datasets, PromptHash demonstrates substantial performance improvements over existing approaches. Notably, on the NUS-WIDE dataset, our method achieves significant gains of 18.22% and 18.65% in image-to-text and text-to-image retrieval tasks, respectively. The code is publicly available at https://github.com/ShiShuMo/PromptHash.

****

V-Stylist: Video Stylization via Collaboration and Reflection of MLLM Agents

Zhengrong Yue, Shaobin Zhuang, Kunchang Li, Yanbo Ding, Yali Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3195-3205

Despite the recent advancement in video stylization, most existing methods struggle to render any video with complex transitions,based on an open style description of user query.To fill this gap,we introduce a generic multi-agent system for video stylization, V-Stylist, by a novel collaboration and reflection paradigm of multi-modal large language models. Specifically, our V-Stylist is a systematical workflow with three key roles: (1) Video Parser decomposes the input video into a number of shots and generates their text prompts of key shot content.Via a concise video-to-shot prompting paradigm,it allows our V-Stylist to effectively handle videos with complex transitions. (2) Style Parser identifies the style in the user query and progressively search the matched style model from a style tree.Via a robust tree-of-thought searching paradigm,it allows our V-Stylist to precisely specify vague style preference in the open user query.(3) Style Artist leverages the matched model to render all the video shots into the required styl

e.Via a novel multi-round self-reflection paradigm,it allows our V-Stylist to ad aptively adjust detail control,according to the style requirement.With such a di stinct design of mimicking human professionals, our V-Stylist achieves a major b reakthrough over the primary challenges for effective and automatic video styliz ation. Moreover, we further construct a new benchmark Text-driven Video Styliza tion Benchmark (TVSBench),which fills the gap to assess stylization of complex v ideos on open user queries. Extensive experiments show that, V-Stylist achieves the state-of-the-art,e.g.,V-Stylist surpasses FRESCO and ControlVideo by 6.05% a nd 4.51% respectively in overall average metrics, marking a significant advance in video stylization.
****

Vision-Language Models Do Not Understand Negation
Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip H.S. Torr, Yoon Kim, Marzyeh Ghassemi; Proceedings of the Computer Vision and Pattern Recog nition Conference (CVPR), 2025, pp. 29612-29622
Many practical vision-language applications require models that understand negat ion, e.g., when using natural language to retrieve images which contain certain objects but not others. Despite advancements in vision-language models (VLMs) th rough large-scale training, their ability to comprehend negation remains underex plored. This study addresses the question: how well do current VLMs understand n egation? We introduce \benchmark , a new benchmark designed to evaluate negatio n understanding across 18 task variations and 79k examples spanning image, video , and medical datasets. The benchmark consists of two core tasks designed to eva luate negation understanding in diverse multimodal settings: Retrieval with Nega tion and Multiple Choice Questions with Negated Captions. Our evaluation reveals that modern VLMs struggle significantly with negation, often performing at chan ce level. To address these shortcomings, we explore a data-centric approach wher ein we finetune CLIP models on large-scale synthetic datasets containing million s of negated captions. We show that this approach can result in a 10% increase in recall on negated queries and a 28% boost in accuracy on multiple-choice ques tions with negated captions.
****

ID-Patch: Robust ID Association for Group Photo Personalization
Yimeng Zhang, Tiancheng Zhi, Jing Liu, Shen Sang, Liming Jiang, Qing Yan, Sijia Liu, Linjie Luo; Proceedings of the Computer Vision and Pattern Recognition Conf erence (CVPR), 2025, pp. 2986-2996
The ability to synthesize personalized group photos and specify the positions of each identity offers immense creative potential. While such imagery can be visu ally appealing, it presents significant challenges for existing technologies. A persistent issue is identity (ID) leakage, where injected facial features interf ere with one another, resulting in low face resemblance, incorrect positioning, and visual artifacts. Existing methods suffer from limitations such as the relia nce on segmentation models, increased runtime, or a high probability of ID leaka ge. To address these challenges, we propose ID-Patch, a novel method that provid es robust association between identities and 2D positions. Our approach generate s an ID patch and ID embeddings from the same facial features: the ID patch is p ositioned on the conditional image for precise spatial control, while the ID emb eddings integrate with text embeddings to ensure high resemblance. Experimental results demonstrate that ID-Patch surpasses baseline methods across metrics, suc h as face ID resemblance, ID-position association accuracy, and generation effic iency. Project Page is: https://byteaigc.github.io/ID-Patch/
****

iG-6DoF: Model-free 6DoF Pose Estimation for Unseen Object via Iterative 3D Gaus sian Splatting
Tuo Cao, Fei Luo, Jiongming Qin, Yu Jiang, Yusen Wang, Chunxia Xiao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 643 6-6446
Traditional methods in pose estimation often rely on precise 3D models or additi onal data such as depth and normals, limiting their generalization, especially w hen objects undergo large translations or rotations. We propose iG-6DoF, a novel

model-free 6D pose estimation method iterative 3D Gaussian Splatting to estimate the pose of unseen objects. We first estimates an initial pose by leveraging multi-scale data augmentation and the rotation-equivariant features to create a better pose hypothesis from a set of candidates. Then, we propose an iterative 3DGS approach through iteratively rendering and comparing the rendered image with the input image to further progressively improve pose estimation accuracy. The proposed end-to-end network consists of an object detector, a multi-scale rotation-equivariant feature based initial pose estimator, and a coarse-to-fine pose refiner. Such combination allows our method to focus on the target object in a complex scene and deal with large movement and weak textures.Such framework allows to deal with large movement and weak texture and complex scene. We conduct extensive experiments on benchmark and real datasets. Our method achieves state-of-the-art results on the LINEMOD, OnePose-LowTexture, and GenMOP datasets, demonstrating its strong generalization to unseen objects and robustness across various scenes.

****

NexusGS: Sparse View Synthesis with Epipolar Depth Priors in 3D Gaussian Splatting

Yulong Zheng, Zicheng Jiang, Shengfeng He, Yandu Sun, Junyu Dong, Huaidong Zhang, Yong Du; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26800-26809

Neural Radiance Field (NeRF) and 3D Gaussian Splatting (3DGS) have noticeably advanced photo-realistic novel view synthesis using images from densely spaced camera viewpoints. However, these methods struggle in few-shot scenarios due to limited supervision. In this paper, we present NexusGS, a 3DGS-based approach that enhances novel view synthesis from sparse-view images by directly embedding depth information into point clouds, without relying on complex manual regularizations. Exploiting the inherent epipolar geometry of 3DGS, our method introduces a novel point cloud densification strategy that initializes 3DGS with a dense point cloud, reducing randomness in point placement while preventing over-smoothing and overfitting. Specifically, NexusGS comprises three key steps: Epipolar Depth Nexus, Flow-Resilient Depth Blending, and Flow-Filtered Depth Pruning. These steps leverage optical flow and camera poses to compute accurate depth maps, while mitigating the inaccuracies often associated with optical flow. By incorporating epipolar depth priors, NexusGS ensures reliable dense point cloud coverage and supports stable 3DGS training under sparse-view conditions. Experiments demonstrate that NexusGS significantly enhances depth accuracy and rendering quality, surpassing state-of-the-art methods by a considerable margin. Furthermore, we validate the superiority of our generated point clouds by substantially boosting the performance of competing methods. Project page: https://usmizuki.github.io/NexusGS/.

****

HyperNet Fields: Efficiently Training Hypernetworks without Ground Truth by Learning Weight Trajectories

Eric Hedlin, Munawar Hayat, Fatih Porikli, Kwang Moo Yi, Shweta Mahajan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22129-22138

To efficiently adapt large models or to train generative models of neural representations, Hypernetworks have drawn interest. While hypernetworks work well, training them is cumbersome, and often requires ground truth optimized weights for each sample. However, obtaining each of these weights is a training problem of its own---one needs to train, e.g., adaptation weights or even an entire neural field for hypernetworks to regress to. In this work, we propose a method to train hypernetworks, without the need for any per-sample ground truth. Our key idea is to learn a Hypernetwork `Field' and estimate the entire trajectory of network weight training instead of simply its converged state. In other words, we introduce an additional input to the Hypernetwork, the convergence state, which then makes it act as a neural field that models the entire convergence pathway of a task network. A critical benefit in doing so is that the gradient of the estimated weights at any convergence state must then match the gradients of the original

task---this constraint alone is sufficient to train the Hypernetwork Field. We d
emonstrate the effectiveness of our method through the task of personalized imag
e generation and 3D shape reconstruction from images and point clouds, demonstra
ting competitive results without any per-sample ground truth.
****

Universal Scene Graph Generation
Shengqiong Wu, Hao Fei, Tat-seng Chua; Proceedings of the Computer Vision and Pa
ttern Recognition Conference (CVPR), 2025, pp. 14158-14168
Scene graph (SG) representations can neatly and efficiently describe scene seman
tics, which has driven sustained intensive research in SG generation. In the rea
l world, multiple modalities often coexist, with different types, such as images
, text, video, and 3D data, expressing distinct characteristics. Unfortunately,
current SG research is largely confined to single-modality scene modeling, preve
nting the full utilization of the complementary strengths of different modality
SG representations in depicting holistic scene semantics. To this end, we introd
uce Universal SG (USG), a novel representation capable of fully characterizing c
omprehensive semantic scenes from any given combination of modality inputs, enco
mpassing modality-invariant and modality-specific scenes. Further, we tailor a n
iche-targeting USG parser, USG-Par, which effectively addresses two key bottlene
cks of cross-modal object alignment and out-of-domain challenges. We design the
USG-Par with modular architecture for end-to-end USG generation, in which we de
vise an object associator to relieve the modality gap for cross-modal object ali
gnment. Further, we propose a text-centric scene contrasting learning mechanism
to mitigate domain imbalances by aligning multimodal objects and relations with
textual SGs. Through extensive experiments, we demonstrate that USG offers a str
onger capability for expressing scene semantics than standalone SGs, and also th
at our USG-Par achieves higher efficacy and performance.
****

RICCARDO: Radar Hit Prediction and Convolution for Camera-Radar 3D Object Detect
ion
Yunfei Long, Abhinav Kumar, Xiaoming Liu, Daniel Morris; Proceedings of the Comp
uter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 22276-22285
Radar hits reflect from points on both the boundary and internal to object outli
nes. This results in a complex distribution of radar hits that depends on factor
s including object category, size and orientation. Current radar-camera fusion m
ethods implicitly account for this with a black-box neural network. In this pape
r, we explicitly utilize a radar hit distribution model to assist fusion. First,
we build a model to predict radar hit distributions conditioned on object prope
rties obtained from a monocular detector. Second, we use the predicted distribut
ion as a kernel to match actual measured radar points in the neighborhood of the
monocular detections, generating matching scores at nearby positions. Finally,
a fusion stage combines context with the kernel detector to refine the matching
scores. Our method achieves the state-of-the-art radar-camera detection performa
nce on nuScenes. Our source code is available at https://github.com/longyunf/ric
cardo.
****

ForestLPR: LiDAR Place Recognition in Forests Attentioning Multiple BEV Density
Images
Yanqing Shen, Turcan Tuna, Marco Hutter, Cesar Cadena, Nanning Zheng; Proceeding
s of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 66
59-6669
Place recognition is essential to maintain global consistency in large-scale loc
alization systems. While research in urban environments has progressed significa
ntly using LiDARs or cameras, applications in natural forest-like environments r
emain largely underexplored. Furthermore, forests present particular challenges
due to high self-similarity and substantial variations in vegetation growth over
time. In this work, we propose a robust LiDAR-based place recognition method fo
r natural forests, ForestLPR. We hypothesize that a set of cross-sectional image
s of the forest's geometry at different heights contains the information needed
to recognize revisiting a place. The cross-sectional images are represented by b

ird's-eye view (BEV) density images of horizontal slices of the point cloud at d
ifferent heights. Our approach utilizes a visual transformer as the shared backb
one to produce sets of local descriptors and introduces a multi-BEV interaction
module to attend to information at different heights adaptively. It is followed
by an aggregation layer that produces a rotation-invariant place descriptor. We
evaluated the efficacy of our method extensively on real-world data from public
benchmarks as well as robotic datasets and compared it against the state-of-the-
art (SOTA) methods. The results indicate that ForestLPR has consistently good pe
rformance on all evaluations and achieves an average increase of 7.38% and 9.11%
 on Recall@1 over the closest competitor on intra-sequence loop closure detectio
n and inter-sequence re-localization, respectively, validating our hypothesis.
****

BLADE: Single-view Body Mesh Estimation through Accurate Depth Estimation
Shengze Wang, Jiefeng Li, Tianye Li, Ye Yuan, Henry Fuchs, Koki Nagano, Shalini
De Mello, Michael Stengel; Proceedings of the Computer Vision and Pattern Recogn
ition Conference (CVPR), 2025, pp. 21991-22000
Single-image human mesh recovery is a challenging task due to the ill-posed natu
re of simultaneous body shape, pose, and camera estimation. Existing estimators
work well on images taken from afar, but they break down as the person moves clo
se to the camera. Moreover, current methods fail to achieve both accurate 3D pos
e and 2D alignment at the same time. Error is mainly introduced by inaccurate pe
rspective projection heuristically derived from orthographic parameters. To reso
lve this long-standing challenge, we present our method BLADE which accurately r
ecovers perspective parameters from a single image without heuristic assumptions
. We start from the inverse relationship between perspective distortion and the
person's Z-translation T_z, and we show that T_z can be reliably estimated from
the image. We then discuss the important role of T_z for accurate human mesh rec
overy estimated from close-range images. Finally, we show that, once T_z and the
 3D human mesh are estimated, one can accurately recover the focal length and fu
ll 3D translation. Extensive experiments on standard benchmarks and real-world c
lose-range images show that our method accurately recovers projection parameters
 from a single image, and consequently attains state-of-the-art accuracy on both
 3D pose estimation and 2D alignment for a wide range of images.
****

AdaMMS: Model Merging for Heterogeneous Multimodal Large Language Models with Un
supervised Coefficient Optimization
Yiyang Du, Xiaochen Wang, Chi Chen, Jiabo Ye, Yiru Wang, Peng Li, Ming Yan, Ji Z
hang, Fei Huang, Zhifang Sui, Maosong Sun, Yang Liu; Proceedings of the Computer
 Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9413-9422
Recently, model merging methods have demonstrated powerful strengths in combinin
g abilities on various tasks from multiple Large Language Models (LLMs). While p
revious model merging methods mainly focus on merging homogeneous models with id
entical architecture, they meet challenges when dealing with Multimodal Large La
nguage Models (MLLMs) with inherent heterogeneous property, including difference
s in model architecture and the asymmetry in the parameter space. In this work,
we propose AdaMMS, a novel model merging method tailored for heterogeneous MLLMs
. Our method tackles the challenges in three steps: mapping, merging and searchi
ng. Specifically, we first design mapping function between models to apply model
 merging on MLLMs with different architecture. Then we apply linear interpolatio
n on model weights to actively adapt the asymmetry in the heterogeneous MLLMs. F
inally in the hyper-parameter searching step, we propose an unsupervised hyper-p
arameter selection method for model merging. As the first model merging method c
apable of merging heterogeneous MLLMs without labeled data, extensive experiment
s on various model combinations demonstrated that AdaMMS outperforms previous mo
del merging methods on various vision-language benchmarks.
****

MoEE: Mixture of Emotion Experts for Audio-Driven Portrait Animation
Huaize Liu, Wenzhang Sun, Donglin Di, Shibo Sun, Jiahui Yang, Changqing Zou, Huj
un Bao; Proceedings of the Computer Vision and Pattern Recognition Conference (C
VPR), 2025, pp. 26222-26231

The generation of talking avatars has achieved significant advancements in precise audio synchronization. However, crafting lifelike talking head videos requires capturing a broad spectrum of emotions and subtle facial expressions. Current methods face fundamental challenges: a) the absence of frameworks for modeling single basic emotional expressions, which restricts the generation of complex emotions such as compound emotions; b) the lack of comprehensive datasets rich in human emotional expressions, which limits the potential of models. To address these challenges, we propose the following innovations: 1) the Mixture of Emotion Experts (MoEE) model, which decouples six fundamental emotions to enable the precise synthesis of both singular and compound emotional states; 2) the DH-FaceEmoVid-150 dataset, specifically curated to include six prevalent human emotional expressions as well as four types of compound emotions, thereby expanding the training potential of emotion-driven models; 3) an emotion-to-latents module that leverages multimodal inputs, aligning diverse control signals--such as audio, text, and labels--to enhance audio-driven emotion control. Through extensive quantitative and qualitative evaluations, we demonstrate that the MoEE framework, in conjunction with the DH-FaceEmoVid-150 dataset, excels in generating complex emotional expressions and nuanced facial details, setting a new benchmark in the field. These datasets will be publicly released.

****

ReCap: Better Gaussian Relighting with Cross-Environment Captures

Jingzhi Li, Zongwei Wu, Eduard Zamfir, Radu Timofte; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21307-21316

Accurate 3D objects relighting in diverse unseen environments is crucial for realistic virtual object placement. Due to the albedo-lighting ambiguity, existing methods often fall short in producing faithful relights. Without proper constraints, observed training views can be explained by numerous combinations of lighting and material attributes, lacking physical correspondence with the actual environment maps used for relighting. In this work, we present ReCap, treating cross-environment captures as multi-task target to provide the missing supervision that cuts through the entanglement. Specifically, ReCap jointly optimizes multiple lighting representations that share a common set of material attributes. This naturally harmonizes a coherent set of lighting representations around the mutual material attributes, exploiting commonalities and differences across varied object appearances. Such coherence enables physically sound lighting reconstruction and robust material estimation -- both essential for accurate relighting. Together with a streamlined shading function and effective post-processing, ReCap outperforms all leading competitors on an expanded relighting benchmark.

****

Split Adaptation for Pre-trained Vision Transformers

Lixu Wang, Bingqi Shang, Yi Li, Payal Mohapatra, Wei Dong, Xiao Wang, Qi Zhu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20092-20102

Vision Transformers (ViTs), extensively pre-trained on large-scale datasets, have become fundamental to foundation models, enabling adaptation to diverse downstream tasks. Existing adaptation methods typically require direct data access, rendering them infeasible in privacy-sensitive domains where clients are often reluctant to share their data. A straightforward solution may be sending the pre-trained ViT to clients for local adaptation, which poses issues of model intellectual property and incurs heavy client computation overhead. To address these issues, we propose a novel split adaptation (SA) method that enables effective downstream adaptation while protecting data and models. SA, inspired by split learning (SL), segments the pre-trained ViT into a frontend and a backend, with only the frontend shared with the client for data representation extraction. But unlike regular SL, SA replaces frontend parameters with low-bit quantized values, preventing direct exposure of the model. SA allows the client to add bi-level noise to the frontend and the extracted data representations, ensuring data protection. Accordingly, SA incorporates data-level and model-level out-of-distribution enhancements to mitigate noise injection's impact. Our SA focuses on the challenging few-shot adaptation and adopts patch retrieval augmentation for overfitting a

lleviation. Extensive experiments on multiple datasets validate SA's superiority over state-of-the-art methods and demonstrate its defense against advanced data reconstruction attacks while preventing model leakage with minimal computation cost on the client side. The source codes can be found at https://github.com/conditionWang/Split_Adaptation.

****

SpatialLLM: A Compound 3D-Informed Design towards Spatially-Intelligent Large Multimodal Models

Wufei Ma, Luoxin Ye, Celso M de Melo, Alan Yuille, Jieneng Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17249-17260

Humans naturally understand 3D spatial relationships, enabling complex reasoning like predicting collisions of vehicles from different directions. Current large multimodal models (LMMs), however, lack of this capability of 3D spatial reasoning. This limitation stems from the scarcity of 3D training data and the bias in current model designs toward 2D data. In this paper, we systematically study the impact of 3D-informed data, architecture, and training setups, introducing SpatialLLM, a large multi-modal model with advanced 3D spatial reasoning abilities. To address data limitations, we develop two types of 3D-informed training data sets: (1) 3D-informed probing data focused on object's 3D location and orientation, and (2) 3D-informed conversation data for complex spatial relationships. Notably, we are the first to curate VQA data that incorporate 3D orientation relationships on real images. Furthermore, we systematically integrate these two types of training data with the architectural and training designs of LMMs, providing a roadmap for optimal design aimed at achieving superior 3D reasoning capabilities. Our SpatialLLM advances machines toward highly capable 3D-informed reasoning, surpassing GPT-4o performance by 8.7%. Our systematic empirical design and the resulting findings offer valuable insights for future research in this direction.

****

SLVR: Super-Light Visual Reconstruction via Blueprint Controllable Convolutions and Exploring Feature Diversity Representation

Ning Ni, Libao Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 400-410

Recently, improving the residual structure and designing efficient convolutions have become important branches of lightweight visual reconstruction model design. We have observed that the feature addition mode (FAM) in existing residual structure tends to lead to slow feature learning or stagnation in feature evolution, a phenomenon we define as network inertia. In addition, although blueprint separable convolutions (BSConv) have proved the dominance of intra-kernel correlation, BSConv forces the blueprint to perform scale transformation on all channels, which may lead to incorrect intra-kernel correlation and introduce useless or disruptive features on some channels and hinder the effective propagation of features. Therefore, in this paper, we rethink the FAM and BSConv for super-light visual reconstruction framework design. First, we design a novel linking mode, called feature diversity evolution link (FDEL), which aims to alleviate the phenomenon of network inertia by reducing the retention of previous low-level features, thereby promoting the evolution of feature diversity. Second, we propose blueprint controllable convolutions (B2Conv). The B2Conv can adaptively pick accurate intra-kernel correlation in the depth-axis, effectively preventing the introduction of useless or disruptive features. Based on FDEL and B2Conv, we develop a super-light super-resolution (SR) framework SLVR for visual reconstruction. Both FDEL and B2Conv can serve as efficient plugins. Extensive experimental results demonstrate the effectiveness of our proposed B2Conv, FDEL, and SLVR. Our code will be available at https://github.com/chongningni/SLVR.

****

Vision-Language Embodiment for Monocular Depth Estimation

Jinchang Zhang, Guoyu Lu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29479-29489

Depth estimation is a core problem in robotic perception and vision tasks, but 3

D reconstruction from a single image presents inherent uncertainties. Current de pth estimation models primarily rely on inter-image relationships for supervised training, often overlooking the intrinsic information provided by the camera it self. We propose a method that embodies the camera model and its physical charac teristics into a deep learning model, computing embodied scene depth through rea l-time interactions with road environments. The model can calculate embodied sce ne depth in real-time based on immediate environmental changes using only the in trinsic properties of the camera, without any additional equipment. By combining embodied scene depth with RGB image features, the model gains a comprehensive p erspective on both geometric and visual details. Additionally, we incorporate te xt descriptions containing environmental content and depth information as priors for scene understanding, enriching the model's perception of objects. This inte gration of image and language -- two inherently ambiguous modalities -- leverage s their complementary strengths for monocular depth estimation. The real-time na ture of the embodied language and depth prior model ensures that the model can c ontinuously adjust its perception and behavior in dynamic environments. Experime ntal results show that the embodied depth estimation method enhances model perfo rmance across different scenes.

****

Layered Image Vectorization via Semantic Simplification

Zhenyu Wang, Jianxi Huang, Zhida Sun, Yuanhao Gong, Daniel Cohen-Or, Min Lu; Pro ceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7728-7738

This work presents a progressive image vectorization technique that reconstructs the raster image as layer-wise vectors from semantic-aligned macro structures t o finer details. Our approach introduces a new image simplification method lever aging the feature-average effect in the Score Distillation Sampling mechanism, a chieving effective visual abstraction from the detailed to coarse. Guided by the sequence of progressive simplified images, we propose a two-stage vectorization process of structural buildup and visual refinement, constructing the vectors i n an organized and manageable manner. The resulting vectors are layered and well -aligned with the target image's explicit and implicit semantic structures. Our method demonstrates high performance across a wide range of images. Comparative analysis with existing vectorization methods highlights our technique's superior ity in creating vectors with high visual fidelity, and more importantly, achievi ng higher semantic alignment and more compact layered representation.

****

Learning Occlusion-Robust Vision Transformers for Real-Time UAV Tracking

You Wu, Xucheng Wang, Xiangyang Yang, Mengyuan Liu, Dan Zeng, Hengzhou Ye, Shuiw ang Li; Proceedings of the Computer Vision and Pattern Recognition Conference (C VPR), 2025, pp. 17103-17113

Single-stream architectures using Vision Transformer (ViT) backbones show great potential for real-time UAV tracking recently. However, frequent occlusions from obstacles like buildings and trees expose a major drawback: these models often lack strategies to handle occlusions effectively. New methods are needed to enha nce the occlusion resilience of single-stream ViT models in aerial tracking. In this work, we propose to learn Occlusion-Robust Representations (ORR) based on V iTs for UAV tracking by enforcing an invariance of the feature representation of a target with respect to random masking operations modeled by a spatial Cox pro cess. Hopefully, this random masking approximately simulates target occlusions, thereby enabling us to learn ViTs that are robust to target occlusion for UAV tr acking. This framework is termed ORTrack. Additionally, to facilitate real-time applications, we propose an Adaptive Feature-Based Knowledge Distillation (AFKD) method to create a more compact tracker, which adaptively mimics the behavior o f the teacher model ORTrack according to the task's difficulty. This student mod el, dubbed ORTrack-D, retains much of ORTrack's performance while offering highe r efficiency. Extensive experiments on multiple benchmarks validate the effectiv eness of our method, demonstrating its state-of-the-art performance. Codes is av ailable at https://github.com/wuyou3474/ORTrack.

****

Plug-and-Play Versatile Compressed Video Enhancement
Huimin Zeng, Jiacheng Li, Zhiwei Xiong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17767-17777
As a widely adopted technique in data transmission, video compression effectively reduces the size of files, making it possible for real-time cloud computing. However, it comes at the cost of visual quality, posing challenges to the robustness of downstream vision models. In this work, we present a versatile codec-aware enhancement framework that reuses codec information to adaptively enhance videos under different compression settings, assisting various downstream vision tasks without introducing computation bottleneck. Specifically, the proposed codec-aware framework consists of a compression-aware adaptation (CAA) network that employs a hierarchical adaptation mechanism to estimate parameters of the frame-wise enhancement network, namely the bitstream-aware enhancement (BAE) network. The BAE network further leverages temporal and spatial priors embedded in the bitstream to effectively improve the quality of compressed input frames. Extensive experimental results demonstrate the superior quality enhancement performance of our framework over existing enhancement methods, as well as its versatility in assisting multiple downstream tasks on compressed videos as a plug-and-play module. Code and models are available at https://huimin-zeng.github.io/PnP-VCVE/.
****
UltraFusion: Ultra High Dynamic Imaging using Exposure Fusion
Zixuan Chen, Yujin Wang, Xin Cai, Zhiyuan You, Zheming Lu, Fan Zhang, Shi Guo, Tianfan Xue; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16111-16121
Capturing high dynamic range (HDR) scenes is one of the most important issues in camera design. Majority of cameras use exposure fusion, which fuses images captured by different exposure levels, to increase dynamic range. However, this approach can only handle images with limited exposure difference, normally 3-4 stops. When applying to very high dynamic range scenes where a large exposure difference is required, this approach often fails due to incorrect alignment or inconsistent lighting between inputs, or tone mapping artifacts. In this work, we propose UltraFusion, the first exposure fusion technique that can merge inputs with 9 stops differences. The key idea is that we model exposure fusion as a guided inpainting problem, where the under-exposed image is used as a guidance to fill the missing information of over-exposed highlights in the over-exposed region. Using an under-exposed image as a soft guidance, instead of a hard constraint, our model is robust to potential alignment issue or lighting variations. Moreover, by utilizing the image prior of the generative model, our model also generates natural tone mapping, even for very high-dynamic range scenes. Our approach outperforms HDR-Transformer on latest HDR benchmarks. Moreover, to test its performance in ultra high dynamic range scenes, we capture a new real-world exposure fusion benchmark, UltraFusion dataset, with exposure differences up to 9 stops, and experiments show that UltraFusion can generate beautiful and high-quality fusion results under various scenarios. Code and data will be available at https://openimaginglab.github.io/UltraFusion.
****
Hearing Anywhere in Any Environment
Xiulong Liu, Anurag Kumar, Paul Calamia, Sebastia V. Amengual, Calvin Murdock, Ishwarya Ananthabhotla, Philip Robinson, Eli Shlizerman, Vamsi Krishna Ithapu, Ruohan Gao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5732-5741
In mixed reality applications, a realistic acoustic experience in spatial environments is as crucial as the visual experience for achieving true immersion. Despite recent advances in neural approaches for Room Impulse Response (RIR) estimation, most existing methods are limited to the single environment on which they are trained, lacking the ability to generalize to new rooms with different geometries and surface materials. We aim to develop a unified model capable of reconstructing the spatial acoustic experience of any environment with minimum additional measurements. To this end, we present xRIR, a framework for cross-room RIR pr

ediction. The core of our generalizable approach lies in combining a geometric f eature extractor, which captures spatial context from panorama depth images, wit h a RIR encoder that extracts detailed acoustic features from only a few referen ce RIR samples. To evaluate our method, we introduce AcousticRooms, a new datase t featuring high-fidelity simulation of over 300,000 RIRs from 260 rooms. Experi ments show that our method strongly outperforms a series of baselines. Furthermo re, we successfully perform sim-to-real transfer by evaluating our model on four real-world environments, demonstrating the generalizability of our approach and the realism of our dataset.
****

Automated Proof of Polynomial Inequalities via Reinforcement Learning
Banglong Liu, Niuniu Qi, Xia Zeng, Lydia Dehbi, Zhengfeng Yang; Proceedings of t he Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5052-506 0
Polynomial inequality proving is fundamental to many mathematical disciplines an d finds wide applications in diverse fields. Current traditional algebraic metho ds are based on searching for a polynomial positive definite representation over a set of basis. However, these methods are limited by truncation degree.To addr ess this issue, this paper proposes an approach based on reinforcement learning to find a Krivine-basis representation for proving polynomial inequalities.Speci fically, we formulate the inequality proving problem as a linear programming (LP) problem and encode it as a basis selection problem using reinforcement learnin g (RL), achieving a non-negative Krivine basis. Moreover, a fast multivariate po lynomial multiplication method based on Fast Fourier Transform (FFT) is employed to enhance the efficiency of action space search. Furthermore, we have implemen ted a tool called APPIRL (Automated Proof of Polynomial Inequalities via Reinfor cement Learning).Experimental evaluation on benchmark problems demonstrates the feasibility and effectiveness of our approach. In addition, APPIRL has been succ essfully applied to solve the maximum stable set problem.
****

Noise-Resistant Video Anomaly Detection via RGB Error-Guided Multiscale Predicti ve Coding and Dynamic Memory
Han Hu, Wenli Du, Peng Liao, Bing Wang, Siyuan Fan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19109-19119
Due to the interference of background noise, existing video anomaly detection me thods are prone to detect some normal events in complex scenes as anomalies. Mea nwhile, we note that the diversity of normal patterns has not been adequately co nsidered, i.e., the normal events that are worthy of reference in the test data have not been properly utilized, which raises the risk of missed and false detec tions. In this work, we combine the tasks of next-frame prediction and predicted -frame reconstruction to propose a noise-resistant video anomaly detection metho d. For the prediction task, we develop an RGB Error-Guided Multiscale Predictive Coding (EG-MPC) framework to overcome the interference of background noise on t he learning of appearance and motion features of objects at various scales, thus achieving high-quality frame prediction. For the reconstruction task, we introd uce the Dynamic Memory Modules (DMMs) into the reconstruction network, and desig n sparse aggregation and selective update strategies for the memory items in the DMMs to effectively represent diverse normal patterns while increasing the diff iculty of reconstructing anomalies, thus making it easier to distinguish between normal and abnormal frames. Extensive experiments on four benchmark datasets de monstrate that our proposed method outperforms state-of-the-art approaches, espe cially in complex scenarios.
****

Frequency Dynamic Convolution for Dense Image Prediction
Linwei Chen, Lin Gu, Liang Li, Chenggang Yan, Ying Fu; Proceedings of the Comput er Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30178-30188
While Dynamic Convolution (DY-Conv) has shown promising performance by enabling adaptive weight selection through multiple parallel weights combined with an att ention mechanism, the frequency response of these weights tends to exhibit high similarity, resulting in high parameter costs but limited adaptability. In this

work, we introduce Frequency Dynamic Convolution (FDConv), a novel approach that mitigates these limitations by learning a fixed parameter budget in the Fourier domain. FDConv divides this budget into frequency-based groups with disjoint Fourier indices, enabling the construction of frequency-diverse weights without increasing the parameter cost. To further enhance adaptability, we propose Kernel Spatial Modulation (KSM) and Frequency Band Modulation (FBM). KSM dynamically adjusts the frequency response of each filter at the spatial level, while FBM decomposes weights into distinct frequency bands in the frequency domain and modulates them dynamically based on local content. Extensive experiments on object detection, segmentation, and classification validate the effectiveness of FDConv. We demonstrate that when applied to ResNet-50, FDConv achieves superior performance with a modest increase of +3.6M parameters, outperforming previous methods that require substantial increases in parameter budgets (e.g., CondConv +90M, KW +76.5M). Moreover, FDConv seamlessly integrates into a variety of architectures, including ConvNeXt, Swin-Transformer, offering a flexible and efficient solution for modern vision tasks. The code is made publicly available at https://github.com/Linwei-Chen/FDConv.

****

IDEA: Inverted Text with Cooperative Deformable Aggregation for Multi-modal Object Re-Identification

Yuhao Wang, Yongfeng Lv, Pingping Zhang, Huchuan Lu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29701-29710

Multi-modal object Re-IDentification (ReID) aims to retrieve specific objects by utilizing complementary information from various modalities. However, existing methods focus on fusing heterogeneous visual features, neglecting the potential benefits of text-based semantic information. To address this issue, we first construct three text-enhanced multi-modal object ReID benchmarks. To be specific, we propose a standardized multi-modal caption generation pipeline for structured and concise text annotations with Multi-modal Large Language Models (MLLMs). Besides, current methods often directly aggregate multi-modal information without selecting representative local features, leading to redundancy and high complexity. To address the above issues, we introduce IDEA, a novel feature learning framework comprising the Inverted Multi-modal Feature Extractor (IMFE) and Cooperative Deformable Aggregation (CDA). The IMFE utilizes Modal Prefixes and an InverseNet to integrate multi-modal information with semantic guidance from inverted text. The CDA adaptively generates sampling positions, enabling the model to focus on the interplay between global features and discriminative local features. With the constructed benchmarks and the proposed modules, our framework can generate more robust multi-modal features under complex scenarios. Extensive experiments on three multi-modal object ReID benchmarks demonstrate the effectiveness of our proposed method.

****

SAM-I2V: Upgrading SAM to Support Promptable Video Segmentation with Less than 0.2% Training Cost

Haiyang Mei, Pengyu Zhang, Mike Zheng Shou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 3417-3426

Foundation models like the Segment Anything Model (SAM) have significantly advanced promptable image segmentation in computer vision. However, extending these capabilities to videos presents substantial challenges, particularly in ensuring precise and temporally consistent mask propagation in dynamic scenes. SAM 2 attempts to address this by training a model on massive image and video data from scratch to learn complex spatiotemporal associations, resulting in huge training costs that hinder research and practical deployment. In this paper, we introduce SAM-I2V, an effective image-to-video upgradation method for cultivating a promptable video segmentation (PVS) model. Our approach strategically upgrades the pre-trained SAM to support PVS, significantly reducing training complexity and resource requirements. To achieve this, we introduce three key innovations: (i) an image-to-video feature extraction upgrader built upon SAM's static image encoder to enable spatiotemporal video perception, (ii) a memory filtering strategy that selects the most relevant past frames for more effective utilization of histori

cal information, and (iii) a memory-as-prompt mechanism leveraging object memory to ensure temporally consistent mask propagation in dynamic scenes. Comprehensive experiments demonstrate that our method achieves over 90% of SAM 2's performance while using only 0.2% of its training cost. Our work presents a resource-efficient pathway to PVS, lowering barriers for further research in PVS model design and enabling broader applications and advancements in the field. Code and model will be available at: https://github.com/showlab/SAM-I2V.
****

GroupMamba: Efficient Group-Based Visual State Space Model
Abdelrahman Shaker, Syed Talal Wasim, Salman Khan, Juergen Gall, Fahad Shahbaz Khan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14912-14922
State-space models (SSMs) have recently shown promise in capturing long-range dependencies with subquadratic computational complexity, making them attractive for various applications. However, purely SSM-based models face critical challenges related to stability and achieving state-of-the-art performance in computer vision tasks. Our paper addresses the challenges of scaling SSM-based models for computer vision, particularly the instability and inefficiency of large model sizes. We introduce a parameter-efficient modulated group mamba layer that divides the input channels into four groups and applies our proposed SSM-based efficient Visual Single Selective Scanning (VSSS) block independently to each group, with each VSSS block scanning in one of the four spatial directions. The Modulated Group Mamba layer also wraps the four VSSS blocks into a channel modulation operator to improve cross-channel communication. Furthermore, we introduce a distillation-based training objective to stabilize the training of large models, leading to consistent performance gains. Our comprehensive experiments demonstrate the merits of the proposed contributions, leading to superior performance over existing methods for image classification on ImageNet-1K, object detection, instance segmentation on MS-COCO, and semantic segmentation on ADE20K. Our tiny variant with 23M parameters achieves state-of-the-art performance with a classification top-1 accuracy of 83.3% on ImageNet-1K, while being 26% efficient in terms of parameters, compared to the best existing Mamba design of same model size. Code and models are available at: https://github.com/Amshaker/GroupMamba
****

How Do I Do That? Synthesizing 3D Hand Motion and Contacts for Everyday Interactions
Aditya Prakash, Benjamin Lundell, Dmitry Andreychuk, David Forsyth, Saurabh Gupta, Harpreet Sawhney; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7026-7036
We tackle the novel problem of predicting 3D hand motion and contact maps (or Interaction Trajectories) given a single RGB view, action text, and a 3D contact point on the object as input. Our approach consists of (1) Interaction Codebook: a VQVAE model to learn a latent codebook of hand poses and contact points, effectively tokenizing interaction trajectories, (2) Interaction Predictor: a transformer-decoder module to predict the interaction trajectory from test time inputs by using an indexer module to retrieve a latent affordance from the learned codebook. To train our model, we develop a data engine that extracts 3D hand poses and contact trajectories from the diverse HoloAssist dataset. We evaluate our model on a benchmark that is 2.5-10X larger than existing works, in terms of diversity of objects and interactions observed, and test for generalization of the model across object categories, action categories, tasks, and scenes. Experimental results show the effectiveness of our approach over transformer & diffusion baselines across all settings.
****

Escaping Plato's Cave: Towards the Alignment of 3D and Text Latent Spaces
Souhail Hadgi, Luca Moschella, Andrea Santilli, Diego Gomez, Qixing Huang, Emanuele Rodolà, Simone Melzi, Maks Ovsjanikov; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19825-19835
Recent works have shown that, when trained at scale, uni-modal 2D vision and text encoders converge to learned features that share remarkable structural propert

ies, despite arising from different representations. However, the role of 3D enc oders with respect to other modalities remains unexplored. Furthermore, existing 3D foundation models that leverage large datasets are typically trained with ex plicit alignment objectives with respect to frozen encoders from other represent ations. In this work, we investigate the possibility of a posteriori alignment o f representations obtained from uni-modal 3D encoders compared to text-based fea ture spaces. We show that naive post-training feature alignment of uni-modal tex t and 3D encoders results in limited performance. We then focus on extracting su bspaces of the corresponding feature spaces and discover that by projecting lear ned representations onto well-chosen lower-dimensional subspaces the quality of alignment becomes significantly higher, leading to improved accuracy on matching and retrieval tasks. Our analysis further sheds light on the nature of these sh ared subspaces, which roughly separate between semantic and geometric data repre sentations. Overall, ours is the first work that helps to establish a baseline f or post-training alignment of 3D uni-modal and text feature spaces, and helps to highlight both the shared and unique properties of 3D data compared to other re presentations.

****

Consistency Posterior Sampling for Diverse Image Synthesis

Vishal Purohit, Matthew Repasky, Jianfeng Lu, Qiang Qiu, Yao Xie, Xiuyuan Cheng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2 025, pp. 28327-28336

Posterior sampling in high-dimensional spaces using generative models holds sign ificant promise for various applications, including but not limited to inverse p roblems and guided generation tasks. Generating diverse posterior samples remain s expensive, as existing methods require restarting the entire generative proces s for each new sample. In this work, we propose a posterior sampling approach th at simulates Langevin dynamics in the noise space of a pre-trained generative mo del. By exploiting the mapping between the noise and data spaces which can be pr ovided by distilled flows or consistency models, our method enables seamless exp loration of the posterior without the need to re-run the full sampling chain, dr astically reducing computational overhead. Theoretically, we prove a guarantee f or the proposed noise-space Langevin dynamics to approximate the posterior, assu ming that the generative model sufficiently approximates the prior distribution. Our framework is experimentally validated on image restoration tasks involving noisy linear and nonlinear forward operators applied to LSUN-Bedroom (256 x 256) and ImageNet (64 x 64) datasets. The results demonstrate that our approach gene rates high-fidelity samples with enhanced semantic diversity even under a limite d number of function evaluations, offering superior efficiency and performance c ompared to existing diffusion-based posterior sampling techniques.

****

IMFine: 3D Inpainting via Geometry-guided Multi-view Refinement

Zhihao Shi, Dong Huo, Yuhongze Zhou, Yan Min, Juwei Lu, Xinxin Zuo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2669 4-26703

Current 3D inpainting and object removal methods are largely limited to front-fa cing scenes, facing substantial challenges when applied to diverse, "unconstrain ed" scenes where the camera orientation and trajectory are unrestricted. To brid ge this gap, we introduce a novel approach that produces inpainted 3D scenes wit h consistent visual quality and coherent underlying geometry across both front-f acing and unconstrained scenes. Specifically, we propose a robust 3D inpainting pipeline that incorporates geometric priors and a multi-view refinement network trained via test-time adaptation, building on a pre-trained image inpainting mod el. Additionally, we develop a novel inpainting mask detection technique to deri ve targeted inpainting masks from object masks, boosting the performance in hand ling unconstrained scenes. To validate the efficacy of our approach, we create a challenging and diverse benchmark that spans a wide range of scenes. Comprehens ive experiments demonstrate that our proposed method substantially outperforms e xisting state-of-the-art approaches.

****

ActiveGAMER: Active GAussian Mapping through Efficient Rendering
Liyan Chen, Huangying Zhan, Kevin Chen, Xiangyu Xu, Qingan Yan, Changjiang Cai, Yi Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16486-16497
We introduce ActiveGAMER, an active mapping system that utilizes 3D Gaussian Splatting (3DGS) to achieve high-quality, real-time scene mapping and exploration. Unlike traditional NeRF-based methods, which are computationally demanding and restrict active mapping performance, our approach leverages the efficient rendering capabilities of 3DGS, allowing effective and efficient exploration in complex environments.The core of our system is a rendering-based information gain module that dynamically identifies the most informative viewpoints for next-best-view planning, enhancing both geometric and photometric reconstruction accuracy. ActiveGAMER also integrates a carefully balanced framework, combining coarse-to-fine exploration, post-refinement, and a global-local keyframe selection strategy to maximize reconstruction completeness and fidelity.Our system autonomously explores and reconstructs environments with state-of-the-art geometric and photometric accuracy and completeness, significantly surpassing existing approaches in both aspects. Extensive evaluations on benchmark datasets such as Replica and MP3D highlight ActiveGAMER's effectiveness in active mapping tasks.
****
DeepCompress-ViT: Rethinking Model Compression to Enhance Efficiency of Vision Transformers at the Edge
Sabbir Ahmed, Abdullah Al Arafat, Deniz Najafi, Akhlak Mahmood, Mamshad Nayeem Rizve, Mohaiminul Al Nahian, Ranyang Zhou, Shaahin Angizi, Adnan Siraj Rakin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30147-30156
Vision Transformers (ViTs) excel in tackling complex vision tasks, yet their substantial size poses significant challenges for applications on resource-constrained edge devices. The increased size of these models leads to higher overhead (e.g., energy, latency) when transmitting model weights between the edge device and the server. Hence, ViTs are not ideal for edge devices where the entire model may not fit on the device. Current model compression techniques often achieve high compression ratios at the expense of performance degradation, particularly for ViTs. To overcome the limitations of existing works, we rethink model compression strategy for ViTs from first principle approach and develop an orthogonal strategy called \method. The objective of the \method is to encode the model weights to a highly compressed encoded representation using a novel training method, denoted as Unified Compression Training (UCT). Proposed UCT is accompanied by a decoding mechanism during inference, which helps to gain any loss of accuracy due to high compression ratio. We further optimize this decoding step by re-ordering the decoding operation using associative property of matrix multiplication, ensuring that the compressed weights can be decoded during inference without incurring any computational overhead. Our extensive experiments across multiple ViT models on modern edge devices show that \method can successfully compress ViTs at high compression ratios (>14x). \method enables the entire model to be stored on edge device, resulting in unprecedented reductions in energy consumption (>1470x) and latency (>68x) for edge ViT inference. Our code is available at \href https://github.com/ML-Security-Research-LAB/DeepCompress-ViT https://github.com/ML-Security-Research-LAB/DeepCompress-ViT .
****
EvOcc: Accurate Semantic Occupancy for Automated Driving Using Evidence Theory
Jonas Kälble, Sascha Wirges, Maxim Tatarchenko, Eddy Ilg; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27467-27476
We present EvOcc, a novel evidential semantic occupancy mapping framework. It consists of two parts: (1) an evidential approach for calculating the ground-truth 3D semantic occupancy maps from noisy LiDAR measurements, and (2) a method for training image-based occupancy estimation models through a new loss formulation. In contrast to state-of-the-art semantic occupancy maps, our approach explicitly models the uncertainty introduced by unobserved spaces or contradicting measurements and we show that using it results in significantly stronger models. Evalu

ated as ray-based mIoU, our evidential semantic occupancy mapping approach impro ves over the baselines by at least 15.8\% for the ground truth and 5.5\% for the trained model. Overall, we make a significant contribution towards more detaile d and uncertainty-aware 3D environment understanding and safe operation in auton omous driving.

****

Positive2Negative: Breaking the Information-Lossy Barrier in Self-Supervised Sin gle Image Denoising

Tong Li, Lizhi Wang, Zhiyuan Xu, Lin Zhu, Wanxuan Lu, Hua Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17924-1 7934

Image denoising enhances image quality, serving as a foundational technique acro ss various computational photography applications. The obstacle to clean image a cquisition in real scenarios necessitates the development of self-supervised ima ge denoising methods only depending on noisy images, especially a single noisy i mage. Existing self-supervised image denoising paradigms (Noise2Noise and Noise2 Void) rely heavily on information-lossy operations, such as downsampling and mas king, culminating in low-quality denoising performance.In this paper, we propose a novel self-supervised single image denoising paradigm, Positive2Negative, to break the information-lossy barrier.Our paradigm involves two key steps: Renoise d Data Construction (RDC) and Denoised Consistency Supervision (DCS). RDC renois es the predicted denoised image by the predicted noise to construct multiple noi sy images, preserving all the information of the original image. DCS ensures con sistency across the multiple denoised images, supervising the network to learn r obust denoising. Our Positive2Negative paradigm achieves state-of-the-art perfor mance in self-supervised single image denoising with significant speed improveme nts. The code is released to the public at https://github.com/Li-Tong-621/P2N.

****

Towards Continual Universal Segmentation

Zihan Lin, Zilei Wang, Xu Wang; Proceedings of the Computer Vision and Pattern R ecognition Conference (CVPR), 2025, pp. 29417-29427

Despite the significant progress in continual image segmentation, existing arts still strive to balance between stability and plasticity. Additionally, they are specialist to specific tasks and models, which hinders the extension to more ge neral situations. In this work, we present CUE, a novel Continual Universal sEgm entation pipeline that not only inherently tackles the stability-plasticity dile mma, but unifies any segmentation across tasks and models as well. Our key insig ht: any segmentation task can be reformulated as an understanding-then-refinemen t paradigm, which is inspired by humans' visual perception system to first perfo rm high-level semantic understanding, then focus on low-level vision cues. We cl aim three desiderata for this design: Continuity by inherently avoiding the stab ility-plasticity dilemma via exploiting the natural differences between high-lev el and low-level knowledge. Generality by unifying and simplifying the landscape towards various segmentation tasks. Efficiency as an interesting by-product by significantly reducing the research effort. Our resulting model, built upon this pipeline by complementary expert models, shows significant improvements over pr evious state-of-the-arts across various segmentation tasks and datasets. We beli eve that our work is a significant step towards making continual segmentation mo re universal and practicable.

****

PGC: Physics-Based Gaussian Cloth from a Single Pose

Michelle Guo, Matt Jen-Yuan Chiang, Igor Santesteban, Nikolaos Sarafianos, Hsiao -yu Chen, Oshri Halimi, Aljaž Boži■, Shunsuke Saito, Jiajun Wu, C. Karen Liu, Tu ur Stuyck, Egor Larionov; Proceedings of the Computer Vision and Pattern Recogni tion Conference (CVPR), 2025, pp. 21215-21225

We introduce a novel approach to reconstruct simulation-ready garments with intr icate appearance. Despite recent advancements, existing methods often struggle t o balance the need for accurate garment reconstruction with the ability to gener alize to new poses and body shapes or require large amounts of data to achieve t his. In contrast, our method only requires a multi-view capture of a single stat

ic frame. We represent garments as hybrid mesh-embedded 3D Gaussian splats, where the Gaussians capture near-field shading and high-frequency details, while the mesh encodes far-field albedo and optimized reflectance parameters. We achieve novel pose generalization by exploiting the mesh from our hybrid approach, enabling physics-based simulation and surface rendering techniques, while also capturing fine details with Gaussians that accurately reconstruct garment details. Our optimized garments can be used for simulating garments on novel poses, and garment relighting. Project page: phys-gaussian-clothing.github.io.
****

Joint Vision-Language Social Bias Removal for CLIP
Haoyu Zhang, Yangyang Guo, Mohan Kankanhalli; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4246-4255
Vision-Language (V-L) pre-trained models such as CLIP show prominent capabilities in various downstream tasks. Despite this promise, V-L models are notoriously limited by their inherent social biases. A typical demonstration is that V-L models often produce biased predictions against specific groups of people, significantly undermining their real-world applicability. Existing approaches endeavor to mitigate the social bias problem in V-L models by removing biased attribute information from model embeddings. However, after our revisiting of these methods, we find that their bias removal is frequently accompanied by greatly compromised V-L alignment capabilities. We then reveal that this performance degradation stems from the unbalanced debiasing in image and text embeddings. To address this issue, we propose a novel V-L debiasing framework to align image and text biases followed by removing them from both modalities. By doing so, our method achieves multi-modal bias mitigation while maintaining the V-L alignment in the debiased embeddings. Additionally, we advocate a new evaluation protocol that can 1) holistically quantify the model debiasing and V-L alignment ability, and 2) evaluate the generalization of social bias removal models. We believe this work will offer new insights and guidance for future studies addressing the social bias problem in CLIP.
****

MV-DUSt3R+: Single-Stage Scene Reconstruction from Sparse Views In 2 Seconds
Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander Schwing, Zhicheng Yan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5283-5293
Recent sparse multi-view scene reconstruction advances like DUSt3R and MASt3R no longer require camera calibration and camera pose estimation. However, they only process a pair of views at a time to infer pixel-aligned pointmaps. When dealing with more than two views, a combinatorial number of error prone pairwise reconstructions are usually followed by an expensive global optimization, which often fails to rectify the pairwise reconstruction errors. To handle more views, reduce errors, and improve inference time, we propose the fast single-stage feed-forward network MV-DUSt3R. At its core are multi-view decoder blocks which exchange information across any number of views while considering one reference view. To make our method robust to reference view selection, we further propose MV-DUSt3R+, which employs cross-reference-view blocks to fuse information across different reference view choices. To further enable novel view synthesis, we extend both by adding and jointly training Gaussian splatting heads. Experiments on multi-view stereo reconstruction, multi-view pose estimation, and novel view synthesis confirm that our methods improve significantly upon prior art. Code released.
****

Explicit Depth-Aware Blurry Video Frame Interpolation Guided by Differential Curves
Zaoming Yan, Pengcheng Lei, Tingting Wang, Faming Fang, Junkang Zhang, Yaomin Huang, Haichuan Song; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1994-2004
Blurry video frame interpolation (BVFI), which aims to generate high-frame-rate clear videos from low-frame-rate blurry inputs, is a challenging yet significant task in computer vision. Current state-of-the-art approaches typically rely on linear or quadratic models to estimate intermediate motion. However, these metho

ds often overlook depth variations that occur during fast object motion, leading to changes in object size and hindering interpolation performance. This paper proposes the Differential Curves-guided Blurry Video Frame Interpolation (DC-BVFI) framework, which leverages the differential curves theory to analyze and mitigate the effects of depth variations caused by object motion. Specifically, DC-BVFI consists of UBNet and MPNet. Unlike prior approaches that rely on optical flow for frame interpolation, MPNet is designed to estimate the 3D scene flow, which facilitates a more precise awareness of depth and velocity variations. Since scene flow cannot be directly inferred in the 2D frame space, UBNet is introduced to transform them into 3D point maps. Extensive experiments demonstrate that the proposed DC-BVFI framework surpasses state-of-the-art performance in simulated and real-world datasets.

****

OFER: Occluded Face Expression Reconstruction

Pratheba Selvaraju, Victoria Fernandez Abrevaya, Timo Bolkart, Rick Akkerman, Tianyu Ding, Faezeh Amjadi, Ilya Zharkov; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26985-26995

Reconstructing 3D face models from a single image is an inherently ill-posed problem, which becomes even more challenging in the presence of occlusions. In addition to fewer available observations, occlusions introduce an extra source of ambiguity where multiple reconstructions can be equally valid. Despite the ubiquity of the problem, very few methods address its multi-hypothesis nature. In this paper we introduce OFER, a novel approach for single-image 3D face reconstruction that can generate plausible, diverse, and expressive 3D faces, even under strong occlusions. Specifically, we train two diffusion models to generate a shape and expression coefficients of face parametric model, conditioned on the input image. This approach captures the multi-modal nature of the problem, generating a distribution of solutions as output. However, to maintain consistency across diverse expressions, the challenge is to select the best matching shape. To achieve this, we propose a novel ranking mechanism that sorts the outputs of the shape diffusion network based on predicted shape accuracy scores. We evaluate our method using standard benchmarks and introduce CO-545, a new protocol and dataset designed to assess the accuracy of expressive faces under occlusion. Our results show improved performance over occlusion-based methods, while also enabling the generation of diverse expressions for a given image.

****

SeaLion: Semantic Part-Aware Latent Point Diffusion Models for 3D Generation

Dekai Zhu, Yan Di, Stefan Gavranovic, Slobodan Ilic; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11789-11798

Denoising diffusion probabilistic models have achieved significant success in point cloud generation, enabling numerous downstream applications, such as generative data augmentation and 3D model editing. However, little attention has been given to generating point clouds with point-wise segmentation labels, as well as to developing evaluation metrics for this task. Therefore, in this paper, we present SeaLion, a novel diffusion model designed to generate high-quality and diverse point cloud with fine-grained segmentation labels. Specifically, we introduce the semantic part-aware latent point diffusion technique, which leverages the intermediate features of the generative models to jointly predict the noise for perturbed latent points and associated part segmentation labels during the denoising process, and subsequently decodes the latent points to point clouds conditioned on part segmentation labels. To effectively evaluate the quality of generated point clouds, we introduce a novel point cloud pairwise distance calculation method named part-aware Chamfer distance (p-CD). This method enables existing metrics, such as 1-NNA, to measure both the local structural quality and inter-part coherence of generated point clouds. Experiments on the large-scale synthetic dataset ShapeNet and real-world medical dataset IntrA, demonstrate that SeaLion achieves remarkable performance in generation quality and diversity, outperforming the existing state-of-the-art model, DiffFacto, by 13.33% and 6.52% on 1-NNA (p-CD) across the two datasets. Experimental analysis shows that SeaLion can be trained semi-supervised, thereby reducing the demand for labeling efforts. Lastl

y, we validate the applicability of SeaLion in generative data augmentation for training segmentation models and the capability of SeaLion to serve as a tool for part-aware 3D shape editing.
****

MonSter: Marry Monodepth to Stereo Unleashes Power
Junda Cheng, Longliang Liu, Gangwei Xu, Xianqi Wang, Zhaoxing Zhang, Yong Deng, Jinliang Zang, Yurui Chen, Zhipeng Cai, Xin Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6273-6282
Stereo matching recovers depth from image correspondences. Existing methods struggle to handle ill-posed regions with limited matching cues, such as occlusions and textureless areas. To address this, we propose MonSter, a novel method that leverages the complementary strengths of monocular depth estimation and stereo matching. MonSter integrates monocular depth and stereo matching into a dual-branch architecture to iteratively improve each other. Confidence-based guidance adaptively selects reliable stereo cues for monodepth scale-shift recovery, and utilizes explicit monocular depth priors to enhance stereo matching at ill-posed regions. Such iterative mutual enhancement enables MonSter to evolve monodepth priors from coarse object-level structures to pixel-level geometry, fully unlocking the potential of stereo matching. As shown in Fig.2, MonSter ranks 1st across five most commonly used leaderboards --- SceneFlow, KITTI 2012, KITTI 2015, Middlebury, and ETH3D. Achieving up to 49.5% improvements over the previous best method (Bad 1.0 on ETH3D). Comprehensive analysis verifies the effectiveness of MonSter in ill-posed regions. In terms of zero-shot generalization, MonSter significantly and consistently outperforms state-of-the-art methods across the board. Code will be released upon acceptance.
****

Toward Real-world BEV Perception: Depth Uncertainty Estimation via Gaussian Splatting
Shu-Wei Lu, Yi-Hsuan Tsai, Yi-Ting Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17124-17133
Bird's-eye view (BEV) perception has gained significant attention because it provides a unified representation to fuse multiple view images and enables a wide range of downstream autonomous driving tasks, such as forecasting and planning. Recent state-of-the-art models utilize projection-based methods which formulate BEV perception as query learning to bypass explicit depth estimation. While we observe promising advancements in this paradigm, they still fall short of real-world applications because of the lack of uncertainty modeling and expensive computational requirement. In this work, we introduce GaussianLSS, an uncertainty-aware BEV perception framework that revisits the unprojection-based method, specifically the Lift-Splat-Shoot (LSS) paradigm, and enhances it with depth uncertainty modeling. Our GaussianLSS represents spatial dispersion by learning a soft depth mean and computing the variance of the depth distribution, which implicitly captures object extents. We then transform the depth distribution into 3D Gaussians and rasterize them to construct uncertainty-aware BEV features. We evaluate GaussianLSS on the nuScenes dataset, achieving state-of-the-art performance compared to unprojection-based methods. In particular, it provides significant advantages in speed, running 2x faster, and in memory efficiency, using 0.3x less memory compared to projection-based methods, while achieving competitive performance with only a 0.7% IoU difference.
****

Efficient Fine-Tuning and Concept Suppression for Pruned Diffusion Models
Reza Shirkavand, Peiran Yu, Shangqian Gao, Gowthami Somepalli, Tom Goldstein, Heng Huang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18619-18629
Recent advances in diffusion generative models have yielded remarkable progress. While the quality of generated content continues to improve, these models have grown considerably in size and complexity. This increasing computational burden poses significant challenges, particularly in resource-constrained deployment scenarios such as mobile devices. The combination of model pruning and knowledge distillation has emerged as a promising solution to reduce computational demands

while preserving generation quality. However, this technique inadvertently propa
gates undesirable behaviors, including the generation of copyrighted content and
 unsafe concepts, even when such instances are absent from the fine-tuning datas
et. In this paper, we propose a novel bilevel optimization framework for pruned
diffusion models that consolidates the fine-tuning and unlearning processes into
 a unified phase. Our approach maintains the principal advantages of distillatio
n--namely, efficient convergence and style transfer capabilities--while selectiv
ely suppressing the generation of unwanted content. This plug-in framework is co
mpatible with various pruning and concept unlearning methods, facilitating effic
ient, safe deployment of diffusion models in controlled environments.
****

Cubify Anything: Scaling Indoor 3D Object Detection
Justin Lazarow, David Griffiths, Gefen Kohavi, Francisco Crespo, Afshin Dehghan;
 Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2
025, pp. 22225-22233
We consider indoor 3D object detection with respect to a single RGB(-D) frame ac
quired from a commodity handheld device. We seek to significantly advance the st
atus quo with respect to both data and modeling. First, we establish that existi
ng datasets have significant limitations to scale, accuracy, and diversity of ob
jects. As a result, we introduce the Cubify-Anything 1M (CA-1M) dataset, which e
xhaustively labels over 400K 3D objects on over 1K highly accurate laser-scanned
 scenes with near-perfect registration to over 3.5K handheld, egocentric capture
s. Next, we establish Cubify Transformer (CuTR), a fully Transformer 3D object d
etection baseline which rather than operating in 3D on point or voxel-based repr
esentations, predicts 3D boxes directly from 2D features derived from RGB(-D) in
puts. While this approach lacks any 3D inductive biases, we show that paired wit
h CA-1M, CuTR outperforms point-based methods, accurately recalling over 62% of
objects in 3D, and is significantly more capable at handling noise and uncertain
ty present in commodity LiDAR-derived depth maps while also providing promising
RGB only performance without architecture changes. Furthermore, by pre-training
on CA-1M, CuTR can outperform point-based methods on a more diverse variant of S
UN RGB-D, supporting the notion that while inductive biases in 3D are useful at
the smaller sizes of existing datasets, they fail to scale to the data-rich regi
me of CA-1M. Overall, this dataset and baseline model provide strong evidence th
at we are moving towards models which can effectively Cubify Anything.
****

WildGS-SLAM: Monocular Gaussian Splatting SLAM in Dynamic Environments
Jianhao Zheng, Zihan Zhu, Valentin Bieri, Marc Pollefeys, Songyou Peng, Iro Arme
ni; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)
, 2025, pp. 11461-11471
We present WildGS-SLAM, a robust and efficient monocular RGB SLAM system designe
d to handle dynamic environments by leveraging uncertainty-aware geometric mappi
ng. Unlike traditional SLAM systems, which assume static scenes, our approach in
tegrates depth and uncertainty information to enhance tracking, mapping, and ren
dering performance in the presence of moving objects. We introduce an uncertaint
y map, predicted by a shallow multi-layer perceptron and DINOv2 features, to gui
de dynamic object removal during both tracking and mapping.This uncertainty map
enhances dense bundle adjustment and Gaussian map optimization, improving recons
truction accuracy. Our system is evaluated on multiple datasets and demonstrates
 artifact-free view synthesis. Results showcase WildGS-SLAM's superior performan
ce in dynamic environments compared to state-of-the-art methods.
****

A Tale of Two Classes: Adapting Supervised Contrastive Learning to Binary Imbala
nced Datasets
David Mildenberger, Paul Hager, Daniel Rueckert, Martin J. Menten; Proceedings o
f the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10305
-10314
Supervised contrastive learning (SupCon) has proven to be a powerful alternative
 to the standard cross-entropy loss for classification of multi-class balanced d
atasets. However, it struggles to learn well-conditioned representations of data

sets with long-tailed class distributions. This problem is potentially exacerbated for binary imbalanced distributions, which are commonly encountered during many real-world problems such as medical diagnosis. In experiments on seven binary datasets of natural and medical images, we show that the performance of SupCon decreases with increasing class imbalance. To substantiate these findings, we introduce two novel metrics that evaluate the quality of the learned representation space. By measuring the class distribution in local neighborhoods, we are able to uncover structural deficiencies of the representation space that classical metrics cannot detect. Informed by these insights, we propose two new supervised contrastive learning strategies tailored to binary imbalanced datasets that improve the structure of the representation space and increase downstream classification accuracy over standard SupCon by up to 35%.
****

DEAL: Data-Efficient Adversarial Learning for High-Quality Infrared Imaging
Zhu Liu, Zijun Wang, Jinyuan Liu, Fanqi Meng, Long Ma, Risheng Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28198-28207
Thermal imaging is often compromised by dynamic, complex degradations caused by hardware limitations and unpredictable environmental factors. The scarcity of high-quality infrared data, coupled with the challenges of dynamic, intricate degradations, makes it difficult to recover details using existing methods. In this paper, we introduce thermal degradation simulation integrated into the training process via a mini-max optimization, by modeling these degraded factors as adversarial attacks on thermal images. The simulation is dynamic to maximize objective functions, thus capturing a broad spectrum of degraded data distributions. This approach enables training with limited data, thereby improving model performance. Additionally, we introduce a dual-interaction network that combines the benefits of spiking neural networks with scale transformation to capture degraded features with sharp spike signal intensities. This architecture ensures compact model parameters while preserving efficient feature representation. Extensive experiments demonstrate that our method not only achieves superior visual quality under diverse single and composited degradation, but also delivers a significant reduction in processing when trained on only fifty clear images, outperforming existing techniques in efficiency and accuracy. The source code will be available at https://github.com/LiuZhu-CV/DEAL.
****

RePerformer: Immersive Human-centric Volumetric Videos from Playback to Photoreal Reperformance
Yuheng Jiang, Zhehao Shen, Chengcheng Guo, Yu Hong, Zhuo Su, Yingliang Zhang, Marc Habermann, Lan Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 11349-11360
Human-centric volumetric videos offer immersive free-viewpoint experiences, yet existing methods focus either on replaying general dynamic scenes or animating human avatars, limiting their ability to re-perform general dynamic scenes. In this paper, we present RePerformer, a novel Gaussian-based representation that unifies playback and re-performance for high-fidelity human-centric volumetric videos. Specifically, we hierarchically disentangle the dynamic scenes into motion Gaussians and appearance Gaussians which are associated in the canonical space. We further employ a Morton-based parameterization to efficiently encode the appearance Gaussians into 2D position and attribute maps. For enhanced generalization, we adopt 2D CNNs to map position maps to attribute maps, which can be assembled into appearance Gaussians for high-fidelity rendering of the dynamic scenes. For re-performance, we develop a semantic-aware alignment module and apply deformation transfer on motion Gaussians, enabling photo-real rendering under novel motions. Extensive experiments validate the robustness and effectiveness of RePerformer, setting a new benchmark for playback-then-reperformance paradigm in human-centric volumetric videos.
****

CheXWorld: Exploring Image World Modeling for Radiograph Representation Learning
Yang Yue, Yulin Wang, Chenxin Tao, Pan Liu, Shiji Song, Gao Huang; Proceedings o

f the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 20778-20788
Humans can develop internal world models that encode common sense knowledge, telling them how the world works and predicting the consequences of their actions. This concept has emerged as a promising direction for establishing general-purpose machine-learning models in recent preliminary works, e.g., for visual representation learning. In this paper, we present CheXWorld, the first effort towards a self-supervised world model for radiographic images. Specifically, our work develops a unified framework that simultaneously models three aspects of medical knowledge essential for qualified radiologists, including 1) local anatomical structures describing the fine-grained characteristics of local tissues (e.g., architectures, shapes, and textures); 2) global anatomical layouts describing the global organization of the human body (e.g., layouts of organs and skeletons); and 3) domain variations that encourage CheXWorld to model the transitions across different appearance domains of radiographs (e.g., varying clarity, contrast, and exposure caused by collecting radiographs from different hospitals, devices, or patients). Empirically, we design tailored qualitative and quantitative analyses, revealing that CheXWorld successfully captures these three dimensions of medical knowledge. Furthermore, transfer learning experiments across eight medical image classification and segmentation benchmarks showcase that CheXWorld significantly outperforms existing SSL methods and large-scale medical foundation models. Code & pre-trained models are available at https://github.com/LeapLabTHU/CheXWorld.
****

Towards Long-Horizon Vision-Language Navigation: Platform, Benchmark and Method
Xinshuai Song, Weixing Chen, Yang Liu, Weikai Chen, Guanbin Li, Liang Lin; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12078-12088
Existing Vision-Language Navigation (VLN) methods primarily focus on single-stage navigation, limiting their effectiveness in multi-stage and long-horizon tasks within complex and dynamic environments. To address these limitations, we propose a novel VLN task, named Long-Horizon Vision-Language Navigation (LH-VLN), which emphasizes long-term planning and decision consistency across consecutive subtasks. Furthermore, to support LH-VLN, we develop an automated data generation platform NavGen, which constructs datasets with complex task structures and improves data utility through a bidirectional, multi-granularity generation approach. To accurately evaluate complex tasks, we construct the Long-Horizon Planning and Reasoning in VLN (LHPR-VLN) benchmark consisting of 3,260 tasks with an average of 150 task steps, serving as the first dataset specifically designed for the long-horizon vision-language navigation task. Furthermore, we propose Independent Success Rate (ISR), Conditional Success Rate (CSR), and CSR weight by Ground Truth (CGT) metrics, to provide fine-grained assessments of task completion. To improve model adaptability in complex tasks, we propose a novel Multi-Granularity Dynamic Memory (MGDM) module that integrates short-term memory blurring with long-term memory retrieval to enable flexible navigation in dynamic environments. Our platform, benchmark and method supply LH-VLN with a robust data generation pipeline, comprehensive model evaluation dataset, reasonable metrics, and a novel VLN model, establishing a foundational framework for advancing LH-VLN.
****

Learning Class Prototypes for Unified Sparse-Supervised 3D Object Detection
Yun Zhu, Le Hui, Hang Yang, Jianjun Qian, Jin Xie, Jian Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9911-9920
Both indoor and outdoor scene perceptions are essential for embodied intelligence. However, current sparse supervised 3D object detection methods focus solely on outdoor scenes without considering indoor settings. To this end, we propose a unified sparse supervised 3D object detection method for both indoor and outdoor scenes through learning class prototypes to effectively utilize unlabeled objects. Specifically, we first propose a prototype-based object mining module that converts the unlabeled object mining into a matching problem between class prototypes and unlabeled features. By using optimal transport matching results, we as

sign prototype labels to high-confidence features, thereby achieving the mining of unlabeled objects. We then present a multi-label cooperative refinement module to effectively recover missed detections in sparse supervised 3D object detection through pseudo label quality control and prototype label cooperation. Experiments show that our method achieves state-of-the-art performance under the one object per scene sparse supervised setting across indoor and outdoor datasets. With only one labeled object per scene, our method achieves about 78%, 90%, and 96% performance compared to the fully supervised detector on ScanNet V2, SUN RGB-D, and KITTI, respectively, highlighting the scalability of our method.

****

BimArt: A Unified Approach for the Synthesis of 3D Bimanual Interaction with Articulated Objects

Wanyue Zhang, Rishabh Dabral, Vladislav Golyanik, Vasileios Choutas, Eduardo Alvarado, Thabo Beeler, Marc Habermann, Christian Theobalt; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27694-27705

We present BimArt, a novel generative approach for synthesizing 3D bimanual hand interactions with articulated objects. Unlike prior works, we do not rely on a reference grasp, a coarse hand trajectory, or separate modes for grasping and articulating. To achieve this, we first generate distance-based contact maps conditioned on the object trajectory with an articulation-aware feature representation, revealing rich bimanual patterns for manipulation. The learned contact prior is then used to guide our hand motion generator, producing diverse and realistic bimanual motions for object movement and articulation. Our work offers key insights into feature representation and contact prior for articulated objects, demonstrating their effectiveness in taming the complex, high-dimensional space of bimanual hand-object interactions. Through comprehensive quantitative experiments, we demonstrate a clear step towards simplified and high-quality hand-object animations that excel over the state-of-the-art in motion quality and diversity.

****

Open-World Amodal Appearance Completion

Jiayang Ao, Yanbei Jiang, Qiuhong Ke, Krista A. Ehinger; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6490-6499

Understanding and reconstructing occluded objects is a challenging problem, especially in open-world scenarios where categories and contexts are diverse and unpredictable. Traditional methods, however, are typically restricted to closed sets of object categories, limiting their use in complex, open-world scenes. We introduce Open-World Amodal Appearance Completion, a training-free framework that expands amodal completion capabilities by accepting flexible text queries as input. Our approach generalizes to arbitrary objects specified by both direct terms and abstract queries. We term this capability reasoning amodal completion, where the system reconstructs the full appearance of the queried object based on the provided image and language query. Our framework unifies segmentation, occlusion analysis, and inpainting to handle complex occlusions and generates completed objects as RGBA elements, enabling seamless integration into applications such as 3D reconstruction and image editing. Extensive evaluations demonstrate the effectiveness of our approach in generalizing to novel objects and occlusions, establishing a new benchmark for amodal completion in open-world settings. Code and datasets available: https://github.com/saraao/amodal.

****

AIGV-Assessor: Benchmarking and Evaluating the Perceptual Quality of Text-to-Video Generation with LMM

Jiarui Wang, Huiyu Duan, Guangtao Zhai, Juntong Wang, Xiongkuo Min; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18869-18880

The rapid advancement of large multimodal models (LMMs) has led to the rapid expansion of artificial intelligence generated videos (AIGVs), which highlights the pressing need for effective video quality assessment (VQA) models designed specifically for AIGVs. Current VQA models generally fall short in accurately assessing the perceptual quality of AIGVs due to the presence of unique distortions, such as unrealistic objects, unnatural movements, or inconsistent visual elements

. To address this challenge, we first present AIGVQA-DB, a large-scale dataset c
omprising 36,576 AIGVs generated by 15 advanced text-to-video models using 1,048
 diverse prompts. With these AIGVs, a systematic annotation pipeline including s
coring and ranking processes is devised, which collects 370k expert ratings to d
ate. Based on AIGVQA-DB, we further introduce AIGV-Assessor, a novel VQA model t
hat leverages spatiotemporal features and LMM frameworks to capture the intricat
e quality attributes of AIGVs, thereby accurately predicting precise video quali
ty scores and video pair preferences. Through comprehensive experiments on both
AIGVQA-DB and existing AIGV databases, AIGV-Assessor demonstrates state-of-the-a
rt performance, significantly surpassing existing scoring or evaluation methods
in terms of multiple perceptual quality dimensions. The dataset and code are rel
eased at https://github.com/IntMeGroup/AIGV-Assessor.
****

CoSpace: Benchmarking Continuous Space Perception Ability for Vision-Language Mo
dels

Yiqi Zhu, Ziyue Wang, Can Zhang, Peng Li, Yang Liu; Proceedings of the Computer
Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29569-29579

Vision-Language Models (VLMs) have recently witnessed significant progress in vi
sual comprehension. As the permitting length of image context grows, VLMs can no
w comprehend a broader range of views and spaces. Current benchmarks provide ins
ightful analysis of VLMs in tasks involving complex visual instructions followin
g, multi-image understanding and spatial reasoning. However, they usually focus
on spatially irrelevant images or discrete images captured from varied viewpoint
s. The compositional characteristic of images captured from a static viewpoint r
emains underestimated. We term this characteristic as Continuous Space Perceptio
n. When observing a scene from a static viewpoint while shifting orientations, i
t produces a series of spatially continuous images, enabling the reconstruction
of the entire space. In this paper, we present CoSpace, a multi-image visual und
erstanding benchmark designed to assess the Continuous Space perception ability
for VLMs. CoSpace contains 2,918 images and 1,626 question-answer pairs, coverin
g seven types of tasks. We conduct evaluation across 19 proprietary and open-sou
rce VLMs. Results reveal that there exist pitfalls on the continuous space perce
ption ability for most of the evaluated models, including proprietary ones. Inte
restingly, we find that the main discrepancy between open-source and proprietary
 models lies not in accuracy but in the consistency of responses. We believe tha
t enhancing the ability of continuous space perception is essential for VLMs to
perform effectively in real-world tasks and encourage further research to advanc
e this capability.
****

Autoregressive Distillation of Diffusion Transformers

Yeongmin Kim, Sotiris Anagnostidis, Yuming Du, Edgar Schönfeld, Jonas Kohler, Ma
rkos Georgopoulos, Albert Pumarola, Ali Thabet, Artsiom Sanakoyeu; Proceedings o
f the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15745
-15756

Diffusion models with transformer architectures have demonstrated promising capa
bilities in generating high-fidelity images and scalability for high resolution.
 However, iterative sampling process required for synthesis is very resource-int
ensive. A line of work has focused on distilling solutions to probability flow
ODEs into few-step student models. Nevertheless, existing methods have been limi
ted by their reliance on the most recent denoised samples as input, rendering th
em susceptible to exposure bias. To address this limitation, we propose AutoRegr
essive Distillation (ARD), a novel approach that leverages the historical trajec
tory of the ODE to predict future steps. ARD offers two key benefits: 1) it miti
gates exposure bias by utilizing a predicted historical trajectory that is less
susceptible to accumulated errors, and 2) it leverages the previous history of t
he ODE trajectory as a more effective source of coarse-grained information. ARD
modifies the teacher transformer architecture by adding token-wise time embeddin
g to mark each input from the trajectory history and employs a block-wise causal
 attention mask for training. Furthermore, incorporating historical inputs only
in lower transformer layers enhances performance and efficiency. We validate the

effectiveness of ARD in a class-conditioned generation on ImageNet and T2I synthesis. Our model achieves a 5xreduction in FID degradation compared to the baseline methods while requiring only 1.1% extra FLOPs on ImageNet-256. Moreover, ARD reaches FID of 1.84 on ImageNet-256 in merely 4 steps and outperforms the publicly available 1024p text-to-image distilled models in prompt adherence score with a minimal drop in FID compared to the teacher. Our code is available at https://github.com/alsdudrla10/ARD.

****

RivuletMLP: An MLP-based Architecture for Efficient Compressed Video Quality Enhancement

Gang He, Weiran Wang, Guancheng Quan, Shihao Wang, Dajiang Zhou, Yunsong Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7342-7352

Quality degradation from video compression manifests both spatially along texture edges and temporally with continuous motion changes. Despite recent advances, extracting aligned spatiotemporal information from adjacent frames remains challenging. This is mainly due to limitations in receptive field size and computational complexity, which makes existing methods struggle to efficiently enhance video quality. To address this issue, we propose RivuletMLP, an MLP-based network architecture. Specifically, our framework first employs a Dynamically Guided Deformable Alignment (DDA) module to adaptively explore and align multi-frame feature information. Subsequently, we introduce two modules for feature reconstruction: a Spatiotemporal Feature Flow (SFF) Module and a Benign Selection Compensation (BSC) module. The SFF module establishes non-local dependencies through an innovative feature permutation mechanism. Additionally, the BSC module utilizes a collaborative strategy of deep feature extraction and local region refinement to alleviate inter frame motion discontinuity caused by compression. Experimental results demonstrate that RivuletMLP achieves superior computational efficiency while maintaining powerful reconstruction capabilities.

****

OmniManip: Towards General Robotic Manipulation via Object-Centric Interaction Primitives as Spatial Constraints

Mingjie Pan, Jiyao Zhang, Tianshu Wu, Yinghao Zhao, Wenlong Gao, Hao Dong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17359-17369

The development of general robotic systems capable of manipulating in unstructured environments is a significant challenge. While Vision-Language Models(VLM) excel in high-level commonsense reasoning, they lack the fine-grained 3D spatial understanding required for precise manipulation tasks. Fine-tuning VLM on robotic datasets to create Vision-Language-Action Models(VLA) is a potential solution, but it is hindered by high data collection costs and generalization issues. To address these challenges, we propose a novel object-centric representation that bridges the gap between VLM's high-level reasoning and the low-level precision required for manipulation. Our key insight is that an object's canonical space, defined by its functional affordances, provides a structured and semantically meaningful way to describe interaction primitives, such as points and directions. These primitives act as a bridge, translating VLM's commonsense reasoning into actionable 3D spatial constraints. In this context, we introduce a dual closed-loop, open-vocabulary robotic manipulation system: one loop for high-level planning through primitive resampling, interaction rendering and VLM checking, and another for low-level execution via 6D pose tracking. This design ensures robust, real-time control without requiring VLM fine-tuning. Extensive experiments demonstrate strong zero-shot generalization across diverse robotic manipulation tasks, highlighting the potential of this approach for automating large-scale data generation.

****

FreeTimeGS: Free Gaussian Primitives at Anytime Anywhere for Dynamic Scene Reconstruction

Yifan Wang, Peishan Yang, Zhen Xu, Jiaming Sun, Zhanhua Zhang, Yong Chen, Hujun Bao, Sida Peng, Xiaowei Zhou; Proceedings of the Computer Vision and Pattern Rec

ognition Conference (CVPR), 2025, pp. 21750-21760
This paper addresses the challenge of reconstructing dynamic 3D scenes with comp lex motions. Some recent works define 3D Gaussian primitives in the canonical sp ace and use deformation fields to map canonical primitives to observation spaces , achieving real-time dynamic view synthesis. However, these methods often strug gle to handle scenes with complex motions due to the difficulty of optimizing de formation fields. To overcome this problem, we propose FreeTimeGS, a novel 4D re presentation that allows Gaussian primitives to appear at arbitrary time and loc ations. In contrast to canonical Gaussian primitives, our representation possess es the strong flexibility, thus improving the ability to model dynamic 3D scenes . In addition, we endow each Gaussian primitive with an motion function, allowin g it to move to neighboring regions over time, which reduces the temporal redund ancy. Experiments results on several datasets show that the rendering quality of our method outperforms recent methods by a large margin. The code will be relea sed for reproducibility.
****

Show and Tell: Visually Explainable Deep Neural Nets via Spatially-Aware Concept Bottleneck Models
Itay Benou, Tammy Riklin Raviv; Proceedings of the Computer Vision and Pattern R ecognition Conference (CVPR), 2025, pp. 30063-30072
Modern deep neural networks have now reached human-level performance across a va riety of tasks. However, unlike humans they lack the ability to explain their de cisions by showing where and telling what concepts guided them. In this work, we present a unified framework for transforming any vision neural network into a s patially and conceptually interpretable model. We introduce a spatially-aware co ncept bottleneck layer that projects "black-box" features of pre-trained backbon e models into interpretable concept maps, without requiring human labels. By tra ining a classification layer over this bottleneck, we obtain a self-explaining m odel that articulates which concepts most influenced its prediction, along with heatmaps that ground them in the input image. Accordingly, we name this method " Spatially-Aware and Label-Free Concept Bottleneck Model" (SALF-CBM). Our results show that the proposed SALF-CBM: (1) Outperforms non-spatial CBM methods, as we ll as the original backbone, on a variety of classification tasks; (2) Produces high-quality spatial explanations, outperforming widely used heatmap-based metho ds on a zero-shot segmentation task; (3) Facilitates model exploration and debug ging, enabling users to query specific image regions and refine the model's deci sions by locally editing its concept maps.
****

DiscoVLA: Discrepancy Reduction in Vision, Language, and Alignment for Parameter -Efficient Video-Text Retrieval
Leqi Shen, Guoqiang Gong, Tianxiang Hao, Tao He, Yifeng Zhang, Pengzhang Liu, Si cheng Zhao, Jungong Han, Guiguang Ding; Proceedings of the Computer Vision and P attern Recognition Conference (CVPR), 2025, pp. 19702-19712
The parameter-efficient adaptation of the image-text pretraining model CLIP for video-text retrieval is a prominent area of research. While CLIP is focused on i mage-level vision-language matching, video-text retrieval demands comprehensive understanding at the video level. Three key discrepancies emerge in the transfer from image-level to video-level: vision, language, and alignment. However, exis ting methods mainly focus on vision while neglecting language and alignment. In this paper, we propose Discrepancy Reduction in Vision, Language, and Alignment (DiscoVLA), which simultaneously mitigates all three discrepancies. Specifically , we introduce Image-Video Features Fusion to integrate image-level and video-le vel features, effectively tackling both vision and language discrepancies. Addit ionally, we generate pseudo image captions to learn fine-grained image-level ali gnment. To mitigate alignment discrepancies, we propose Image-to-Video Alignment Distillation, which leverages image-level alignment knowledge to enhance video- level alignment. Extensive experiments demonstrate the superiority of our DiscoV LA. In particular, on MSRVTT with CLIP (ViT-B/16), DiscoVLA outperforms previous methods by 2.2% R@1 and 7.5% R@sum. The code is available at https://github.com /LunarShen/DsicoVLA.

****
Reanimating Images using Neural Representations of Dynamic Stimuli
Jacob Yeung, Andrew F. Luo, Gabriel Sarch, Margaret M. Henderson, Deva Ramanan, Michael J. Tarr; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 5331-5343
While computer vision models have made incredible strides in static image recognition, they still do not match human performance in tasks that require the understanding of complex, dynamic motion. This is notably true for real-world scenarios where embodied agents face complex and motion-rich environments. Our approach, BrainNRDS (Brain-Neural Representations of Dynamic Stimuli), leverages state-of-the-art video diffusion models to decouple static image representation from motion generation, enabling us to utilize fMRI brain activity for a deeper understanding of human responses to dynamic visual stimuli. Conversely, we also demonstrate that information about the brain's representation of motion can enhance the prediction of optical flow in artificial systems. Our novel approach leads to four main findings: (1) Visual motion, represented as fine-grained, object-level resolution optical flow, can be decoded from brain activity generated by participants viewing video stimuli; (2) Video encoders outperform image-based models in predicting video-driven brain activity; (3) Brain-decoded motion signals enable realistic video reanimation based only on the initial frame of the video; and (4) We extend prior work to achieve full video decoding from video-driven brain activity. BrainNRDS advances our understanding of how the brain represents spatial and temporal information in dynamic visual scenes. Our findings demonstrate the potential of combining brain imaging with video diffusion models for developing more robust and biologically-inspired computer vision systems. We show additional decoding and encoding examples on this site: https://brain-nrds.github.io/.
****
Visual-Instructed Degradation Diffusion for All-in-One Image Restoration
Wenyang Luo, Haina Qin, Zewen Chen, Libin Wang, Dandan Zheng, Yuming Li, Yufan Liu, Bing Li, Weiming Hu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12764-12777
Image restoration tasks like deblurring, denoising, and dehazing usually need distinct models for each degradation type, restricting their generalization in real-world scenarios with mixed or unknown degradations. In this work, we propose Defusion, a novel all-in-one image restoration framework that utilizes visual instruction-guided degradation diffusion. Unlike existing methods that rely on task-specific models or ambiguous text-based priors, Defusion constructs explicit visual instructions that align with the visual degradation patterns. These instructions are grounded by applying degradations to standardized visual elements, capturing intrinsic degradation features while agnostic to image semantics. Defusion then uses these visual instructions to guide a diffusion-based model that operates directly in the degradation space, where it reconstructs high-quality images by denoising the degradation effects with enhanced stability and generalizability. Comprehensive experiments demonstrate that Defusion outperforms state-of-the-art methods across diverse image restoration tasks, including complex and real-world degradations.
****
Insightful Instance Features for 3D Instance Segmentation
Wonseok Roh, Hwanhee Jung, Giljoo Nam, Dong In Lee, Hyeongcheol Park, Sang Ho Yoon, Jungseock Joo, Sangpil Kim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14057-14067
Recent 3D Instance Segmentation methods typically encode hundreds of instance-wise candidates with instance-specific information in various ways and refine them into final masks. However, they have yet to fully explore the benefit of these candidates. They overlook the valuable cues encoded in multiple candidates that represent different parts of the same instance, resulting in fragments. Also, they often fail to capture the precise spatial range of 3D instances, primarily due to inherent noises from sparse and unordered point clouds. In this work, to address these challenges, we propose IKNE, a novel instance-wise knowledge enhancement approach. We first introduce an Instance-wise Knowledge Aggregation (IKA) t

o associate scattered single instance details by optimizing correlations among c
andidates representing the same instance. Moreover, we present an Instance-wise
Structural Guidance (ISG) to enhance the spatial understanding of candidates usi
ng structural cues from ambiguity-reduced features. Here, we utilize a simple ye
t effective truncated singular value decomposition algorithm to minimize inheren
t noises of 3D features. In our extensive experiments on large-scale datasets, S
canNetV2, ScanNet200, S3DIS, and STPLS3D, IKNE outperforms existing works. We va
lidate the effectiveness of our modules in both kernel-based and transformer-bas
ed architectures.
****

Learning 4D Panoptic Scene Graph Generation from Rich 2D Visual Scene
Shengqiong Wu, Hao Fei, Jingkang Yang, Xiangtai Li, Juncheng Li, Hanwang Zhang,
Tat-seng Chua; Proceedings of the Computer Vision and Pattern Recognition Confer
ence (CVPR), 2025, pp. 24539-24549
The latest emerged 4D Panoptic Scene Graph (4D-PSG) provides an advanced-ever re
presentation for comprehensively modeling the dynamic 4D visual real world. Unfo
rtunately, current pioneering 4D-PSG research can largely suffer from data scarc
ity issues severely, as well as the resulting out-of-vocabulary problems; also,
the pipeline nature of the benchmark generation method can lead to suboptimal pe
rformance. To address these challenges, this paper investigates a novel framewor
k for 4D-PSG generation that leverages rich 2D visual scene annotations to enhan
ce 4D scene learning. First, we introduce a 4D Large Language Model (4D-LLM) int
egrated with a 3D mask decoder for end-to-end generation of 4D-PSG. A chained SG
 inference mechanism is further designed to exploit LLMs' open-vocabulary capabi
lities to infer accurate and comprehensive object and relation labels iterativel
y. Most importantly, we propose a 2D-to-4D visual scene transfer learning framew
ork, where a spatial-temporal scene transcending strategy effectively transfers
dimension-invariant features from abundant 2D SG annotations to 4D scenes, effec
tively compensating for data scarcity in 4D-PSG. Extensive experiments on the be
nchmark data demonstrate that we strikingly outperform baseline models by a larg
e margin, highlighting the effectiveness of our method.
****

Knowledge Bridger: Towards Training-Free Missing Modality Completion
Guanzhou Ke, Shengfeng He, Xiaoli Wang, Bo Wang, Guoqing Chao, Yuanyang Zhang, Y
i Xie, Hexing Su; Proceedings of the Computer Vision and Pattern Recognition Con
ference (CVPR), 2025, pp. 25864-25873
Previous successful approaches to missing modality completion rely on carefully
designed fusion techniques and extensive pre-training on complete data, which ca
n limit their generalizability in out-of-domain (OOD) scenarios. In this study,
we pose a new challenge: can we develop a missing modality completion model that
 is both resource-efficient and robust to OOD generalization? To address this, w
e present a training-free framework for missing modality completion that leverag
es large multimodal models (LMMs). Our approach, termed the "Knowledge Bridger",
 is modality-agnostic and integrates generation and ranking of missing modalitie
s. By defining domain-specific priors, our method automatically extracts structu
red information from available modalities to construct knowledge graphs. These e
xtracted graphs connect the missing modality generation and ranking modules thro
ugh the LMM, resulting in high-quality imputations of missing modalities. Experi
mental results across both general and medical domains show that our approach co
nsistently outperforms competing methods, including in OOD generalization. Addit
ionally, our knowledge-driven generation and ranking techniques demonstrate supe
riority over variants that directly employ LMMs for generation and ranking, offe
ring insights that may be valuable for applications in other domains.
****

EmoDubber: Towards High Quality and Emotion Controllable Movie Dubbing
Gaoxiang Cong, Jiadong Pan, Liang Li, Yuankai Qi, Yuxin Peng, Anton van den Heng
el, Jian Yang, Qingming Huang; Proceedings of the Computer Vision and Pattern Re
cognition Conference (CVPR), 2025, pp. 15863-15873
Given a piece of text, a video clip, and a reference audio, the movie dubbing ta
sk aims to generate speech that aligns with the video while cloning the desired

voice. The existing methods have two primary deficiencies: (1) They struggle to simultaneously hold audio-visual sync and achieve clear pronunciation; (2) They lack the capacity to express user-defined emotions. To address these problems, we propose EmoDubber, an emotion-controllable dubbing architecture that allows users to specify emotion type and emotional intensity while satisfying high-quality lip sync and pronunciation. Specifically, we first design Lip-related Prosody Aligning (LPA), which focuses on learning the inherent consistency between lip motion and prosody variation by duration level contrastive learning to incorporate reasonable alignment. Then, we design Pronunciation Enhancing (PE) strategy to fuse the video-level phoneme sequences by efficient conformer to improve speech intelligibility. Next, the speaker identity adapting module decodes acoustics prior and inject the speaker style embedding. After that, the proposed Flow-based User Emotion Controlling (FUEC) is used to synthesize waveform by flow matching prediction network conditioned on acoustics prior. In this process, the FUEC determines the gradient direction and guidance scale based on the user's emotion instructions by the positive and negative guidance mechanism, which focuses on amplifying the desired emotion while suppressing others. Extensive experimental results demonstrate favorable performance compared to several state-of-the-art methods. The code and trained models will be made available at https://github.com/GalaxyCong/DubFlow.

****

DepthCrafter: Generating Consistent Long Depth Sequences for Open-world Videos

Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, Ying Shan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2005-2015

Estimating video depth in open-world scenarios is challenging due to the diversity of videos in appearance, content motion, camera movement, and length. We present DepthCrafter, an innovative method for generating temporally consistent long depth sequences with intricate details for open-world videos, without requiring any supplementary information such as camera poses or optical flow. The generalization ability to open-world videos is achieved by training the video-to-depth model from a pre-trained image-to-video diffusion model, through our meticulously designed three-stage training strategy. Our training approach enables the model to generate depth sequences with variable lengths at one time, up to 110 frames, and harvest both precise depth details and rich content diversity from realistic and synthetic datasets. We also propose an inference strategy that can process extremely long videos through segment-wise estimation and seamless stitching. Comprehensive evaluations on multiple datasets reveal that DepthCrafter achieves state-of-the-art performance in open-world video depth estimation under zero-shot settings. Furthermore, DepthCrafter facilitates various downstream applications, including depth-based visual effects and conditional video generation.

****

TexGarment: Consistent Garment UV Texture Generation via Efficient 3D Structure-Guided Diffusion Transformer

Jialun Liu, Jinbo Wu, Xiaobo Gao, Jiakui Hu, Bojun Xiong, Xing Liu, Chen Zhao, Hongbin Pei, Haocheng Feng, Yingying Li, Errui Ding, Jingdong Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26566-26575

This paper introduces TexGarment, an efficient method for synthesizing high-quality, 3D-consistent garment textures in UV space. Traditional approaches based on 2D-to-3D mapping often suffer from 3D inconsistency, while methods learning from limited 3D data lack sufficient texture diversity. These limitations are particularly problematic in garment texture generation, where high demands exist for both detail and variety. To address these challenges, TexGarment leverages a pre-trained text-to-image diffusion Transformer model with robust generalization capabilities, introducing structural information to guide the model in generating 3D-consistent garment textures in a single inference step. Specifically, We utilize the 2D UV position map to guide the layout during the UV texture generation process, ensuring a coherent texture arrangement and enhancing it by integrating global 3D structural information from the mesh surface point cloud. This combin

ed guidance effectively aligns 3D structural integrity with 2D layout. Our metho d efficiently generates high-quality, diverse UV textures in a single inference step while maintaining 3D consistency. Experimental results validate the effecti veness of TexGarment, achieving state-of-the-art performance in 3D garment textu re generation.
****

A Hubness Perspective on Representation Learning for Graph-Based Multi-View Clus tering
Zheming Xu, He Liu, Congyan Lang, Tao Wang, Yidong Li, Michael C. Kampffmeyer; P roceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 202 5, pp. 15528-15537
Recent graph-based multi-view clustering (GMVC) methods typically encode view fe atures into high-dimensional spaces and construct graphs based on distance simil arity. However, the high dimensionality of the embeddings often leads to the hub ness problem, where a few points repeatedly appear in the nearest neighbor lists of other points. We show that this negatively impacts the extracted graph struc tures and message passing, thus degrading clustering performance. To the best of our knowledge, we are the first to highlight the detrimental effect of hubness in GMVC methods and introduce the hubREP (hub-aware Representation Embedding and Pairing) framework. Specifically, we propose a simple yet effective encoder tha t reduces hubness while preserving neighborhood topology within each view. Addit ionally, we propose a hub-aware pairing module to maintain structure consistency across views, efficiently enhancing the view-specific representations. The prop osed hubREP is lightweight compared to the conventional autoencoders used in sta te-of-the-art GMVC methods and can be integrated into existing GMVC methods that mostly focus on novel fusion mechanisms, further boosting their performance. Co mprehensive experiments performed on eight benchmarks confirm the superiority of our method. The code is available at https://github.com/zmxu196/hubREP.
****

Spatial-Temporal Graph Diffusion Policy with Kinematic Modeling for Bimanual Rob otic Manipulation
Qi Lv, Hao Li, Xiang Deng, Rui Shao, Yinchuan Li, Jianye Hao, Longxiang Gao, Mic hael Yu Wang, Liqiang Nie; Proceedings of the Computer Vision and Pattern Recogn ition Conference (CVPR), 2025, pp. 17394-17404
Despite the significant success of imitation learning in robotic manipulation, i ts application to bimanual tasks remains highly challenging. Existing approaches mainly learn a policy to predict a distant next-best end-effector pose (NBP) an d then compute the corresponding joint rotation angles for motion using inverse kinematics. However, they suffer from two important issues: (1) rarely consideri ng the physical robotic structure , which may cause self-collisions or interfere nces, and (2) overlooking the kinematics constraint , which may result in the pr edicted poses not conforming to the actual limitations of the robot joints. In t his paper, we propose Kinematics enhanced Spatial-TemporAl gRaph Diffuser (KStar Diffuser). Specifically, (1)to incorporate the physical robot structure informa tion into action prediction, KStar Diffuser maintains a dynamic spatial-temporal graph according to the physical bimanual joint motions at continuous timesteps. This dynamic graph serves as the robot-structure condition for denoising the ac tions; (2) to make the NBP learning objective consistent with kinematics, we int roduce the differentiable kinematics to provide the reference for optimizing KSt ar Diffuser. This module regularizes the policy to predict more reliable and kin ematics-aware next end-effector poses. Experimental results show that our method effectively leverages the physical structural information and generates kinemat ics-aware actions in both simulation and real-world.
****

Semi-Supervised State-Space Model with Dynamic Stacking Filter for Real-World Vi deo Deraining
Shangquan Sun, Wenqi Ren, Juxiang Zhou, Shu Wang, Jianhou Gan, Xiaochun Cao; Pro ceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26114-26124
Significant progress has been made in video restoration under rainy conditions o

ver the past decade, largely propelled by advancements in deep learning. Neverth eless, existing methods that depend on paired data struggle to generalize effect ively to real-world scenarios, primarily due to the disparity between synthetic and authentic rain effects. To address these limitations, we propose a dual-bran ch spatio-temporal state-space model to enhance rain streak removal in video seq uences. Specifically, we design spatial and temporal state-space model layers to extract spatial features and incorporate temporal dependencies across frames, r espectively. To improve multi-frame feature fusion, we derive a dynamic stacking filter, which adaptively approximates statistical filters for superior pixel-wi se feature refinement. Moreover, we integrate a median stacking loss to enable s emi-supervised learning by generating pseudo-clean patches based on the sparsity prior of rain. To further explore the capacity of deraining models in supportin g other vision-based tasks in rainy environments, we introduce a novel real-worl d benchmark focused on object detection and tracking in rainy conditions. Our me thod is extensively evaluated across multiple benchmarks containing numerous syn thetic and real-world rainy videos, consistently demonstrating its superiority i n quantitative metrics, visual quality, efficiency, and its utility for downstre am tasks.

****

Rethinking Vision-Language Model in Face Forensics: Multi-Modal Interpretable Fo rged Face Detector

Xiao Guo, Xiufeng Song, Yue Zhang, Xiaohong Liu, Xiaoming Liu; Proceedings of th e Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 105-116

Deepfake detection is a long-established research topic vital for mitigating the spread of malicious misinformation. Unlike prior methods that provide either bi nary classification results or textual explanations separately, we introduce a n ovel method capable of generating both simultaneously. Our method harnesses the multi-modal learning capability of the pre-trained CLIP and the unprecedented in terpretability of large language models (LLMs) to enhance both the generalizatio n and explainability of deepfake detection. Specifically, we introduce a multi-m odal face forgery detector (M2F2-Det) that employs tailored face forgery prompt learning, incorporating the pre-trained CLIP to improve generalization to unseen forgeries. Also, M2F2-Det incorporates an LLM to provide detailed textual expla nations of its detection decisions, enhancing interpretability by bridging the g ap between natural language and subtle cues of facial forgeries. Empirically, we evaluate M2F2-Det on both detection and explanation generation tasks, where it achieves state-of-the-art performance, demonstrating its effectiveness in identi fying and explaining diverse forgeries. Source code and models are available at this link.

****

TIDE: Training Locally Interpretable Domain Generalization Models Enables Test-t ime Correction

Aishwarya Agarwal, Srikrishna Karanam, Vineet Gandhi; Proceedings of the Compute r Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30210-30220

We consider the problem of single-source domain generalization. Existing methods typically rely on extensive augmentations to synthetically cover diverse domain s during training. However, they struggle with semantic shifts (e.g., background and viewpoint changes), as they often learn global features instead of local co ncepts that tend to be domain invariant. To address this gap, we propose an appr oach that compels models to leverage such local concepts during prediction. Give n no suitable dataset with per-class concepts and localization maps exists, we f irst develop a novel pipeline to generate annotations by exploiting the rich fea tures of diffusion and large-language models. Our next innovation is TIDE, a nov el training scheme with a concept saliency alignment loss that ensures model foc us on the right per-concept regions and a local concept contrastive loss that pr omotes learning domain-invariant concept representations. This not only gives a robust model but also can be visually interpreted using the predicted concept sa liency maps. Given these maps at test time, our final contribution is a new corr ection algorithm that uses the corresponding local concept representations to it eratively refine the prediction until it aligns with prototypical concept repres

entations that we store at the end of model training. We evaluate our approach e
xtensively on four standard DG benchmark datasets and substantially outperform t
he current state-of-the-art (12% improvement on average) while also demonstratin
g that our predictions can be visually interpreted.
****

VSNet: Focusing on the Linguistic Characteristics of Sign Language
Yuhao Li, Xinyue Chen, Hongkai Li, Xiaorong Pu, Peng Jin, Yazhou Ren; Proceeding
s of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24
320-24330
Sign language is a visual language expressed through complex movements of the up
per body. The human skeleton plays a critical role in sign language recognition
due to its good separation from the video background. However, mainstream skelet
on-based sign language recognition models often overly focus on the natural conn
ections between joints, treating sign language as ordinary human movements, whic
h neglects its linguistic characteristics. We believe that just as letters form
words, each sign language gloss can also be decomposed into smaller visual symbo
ls. To fully harness the potential of skeleton data, this paper proposes a novel
 joint fusion strategy and a visual symbol attention model. Specifically, we fir
st input the complete set of skeletal joints, and after dynamically exchanging j
oint information, we discard the parts with the weakest connections to other joi
nts, resulting in a fused, simplified skeleton. Then, we group the joints most l
ikely to express the same visual symbol and discuss the joint movements within e
ach group separately. To validate the superiority of our method, we conduct exte
nsive experiments on multiple public benchmark datasets. The results show that,
without complex pre-training, we still achieve new state-of-the-art performance.
****

Active Hyperspectral Imaging Using an Event Camera
Bohan Yu, Jinxiu Liang, Zhuofeng Wang, Bin Fan, Art Subpa-asa, Boxin Shi, Imari
Sato; Proceedings of the Computer Vision and Pattern Recognition Conference (CVP
R), 2025, pp. 929-939
Hyperspectral imaging plays a critical role in numerous scientific and industria
l fields. Conventional hyperspectral imaging systems often struggle with the tra
de-off between capture speed, spectral resolution, and bandwidth, particularly i
n dynamic environments. In this work, we present a novel event-based active hype
rspectral imaging system designed for real-time capture with low bandwidth in dy
namic scenes. By combining an event camera with a dynamic illumination strategy,
 our system achieves unprecedented temporal resolution while maintaining high sp
ectral fidelity, all at a fraction of the bandwidth requirements of traditional
systems. Unlike basis-based methods that sacrifice spectral resolution for effic
iency, our approach enables continuous spectral sampling through an innovative "
sweeping rainbow" illumination pattern synchronized with a rotating mirror array
. The key insight is leveraging the sparse, asynchronous nature of event cameras
 to encode spectral variations as temporal contrasts, effectively transforming t
he spectral reconstruction problem into a series of geometric constraints. Exten
sive evaluations of both synthetic and real data demonstrate that our system out
performs state-of-the-art methods in temporal resolution while maintaining compe
titive spectral reconstruction quality.
****

Bridging the Gap between Gaussian Diffusion Models and Universal Quantization fo
r Image Compression
Lucas Relic, Roberto Azevedo, Yang Zhang, Markus Gross, Christopher Schroers; Pr
oceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025
, pp. 2449-2458
Generative neural image compression supports data representation with extremely
low bitrate, allowing clients to synthesize details and consistently producing h
ighly realistic images. By leveraging the similarities between quantization erro
r and additive noise, diffusion-based generative image compression codecs can be
 built using a latent diffusion model to "denoise" the artifacts introduced by q
uantization. However, we identify three critical gaps in previous approaches fol
lowing this paradigm (namely, the noise level, noise type, and discretization ga

ps) that result in the quantized data falling out of the data distribution known by the diffusion model. In this work, we propose a novel quantization-based forward diffusion process with theoretical foundations that tackles all three aforementioned gaps. We achieve this through universal quantization with a carefully tailored quantization schedule and a diffusion model trained with uniform noise. Compared to previous work, our proposal produces consistently realistic and detailed reconstructions, even at very low bitrates. In such a regime, we achieve the best rate-distortion-realism performance, outperforming previous related works.

****

ZeroVO: Visual Odometry with Minimal Assumptions
Lei Lai, Zekai Yin, Eshed Ohn-Bar; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17092-17102
We introduce ZeroVO, a novel visual odometry (VO) algorithm that achieves zero-shot generalization across diverse cameras and environments, overcoming limitations in existing methods that depend on predefined or static camera calibration setups. Our approach incorporates three main innovations. First, we design a calibration-free, geometry-aware network structure capable of handling noise in estimated depth and camera parameters. Second, we introduce a language-based prior that infuses semantic information to enhance robust feature extraction and generalization to previously unseen domains. Third, we develop a flexible, semi-supervised training paradigm that iteratively adapts to new scenes using unlabeled data, further boosting the models' ability to generalize across diverse real-world scenarios. We analyze complex autonomous driving contexts, demonstrating over 30% improvement against prior methods on three standard benchmarks--KITTI, nuScenes, and Argoverse 2--as well as a newly introduced, high-fidelity synthetic dataset derived from Grand Theft Auto (GTA). By not requiring fine-tuning or camera calibration, our work broadens the applicability of VO, providing a versatile solution for real-world deployment at scale.

****

VideoRefer Suite: Advancing Spatial-Temporal Object Understanding with Video LLM
Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, Jianke Zhu, Lidong Bing; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18970-18980
Video Large Language Models (Video LLMs) have recently exhibited remarkable capabilities in general video understanding.However, they mainly focus on holistic comprehension and struggle with capturing fine-grained spatial and temporal details. Besides, the lack of high-quality object-level video instruction data and a comprehensive benchmark further hinders their advancements. To tackle these challenges, we introduce the VideoRefer Suite to empower Video LLM for finer-level spatial-temporal video understanding, i.e., enabling perception and reasoning on any objects throughout the video. Specially, we thoroughly develop VideoRefer Suite across three essential aspects: dataset, model, and benchmark. Firstly, we introduce a multi-agent data engine to meticulously curate a large-scale, high-quality object-level video instruction dataset, termed VideoRefer-700K. Next, we present the VideoRefer model, which equips a versatile spatial-temporal object encoder to capture precise regional and sequential representations. Finally, we meticulously create a VideoRefer-Bench to comprehensively assess the spatial-temporal understanding capability of a Video LLM, evaluating it across various aspects. Extensive experiments and analyses demonstrate that our VideoRefer model not only achieves promising performance on video referring benchmarks but also facilitates general video understanding capabilities. Our codes, models, and dataset will be made publicly available.

****

Learning to Sample Effective and Diverse Prompts for Text-to-Image Generation
Taeyoung Yun, Dinghuai Zhang, Jinkyoo Park, Ling Pan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23625-23635
Recent advances in text-to-image diffusion models have demonstrated impressive image generation capabilities. However, it remains challenging to control the gen

eration process with desired properties (e.g., aesthetic quality, user intention), which can be expressed as black-box reward functions. Recent advances in text -to-image diffusion models have achieved impressive image generation capabilitie s. However, it remains challenging to control the generation process with desire d properties (e.g., aesthetic quality, user intention), which can be expressed a s black-box reward functions. In this paper, we focus on prompt adaptation, whic h refines the original prompt into model-preferred prompts to generate desired i mages. While prior work uses reinforcement learning (RL) to optimize prompts, we observe that applying RL often results in generating similar postfixes and dete rministic behaviors. To this end, we introduce Prompt Adaptation with GFlowNets (PAG), a novel approach that frames prompt adaptation as a probabilistic inferen ce problem. Our key insight is that leveraging Generative Flow Networks (GFlowNe ts) allows us to shift from reward maximization to sampling from an unnormalized density function, enabling both high-quality and diverse prompt generation. How ever, we identify that a naive application of GFlowNets suffers from mode collap se and uncovers a previously overlooked phenomenon: the progressive loss of neur al plasticity in the model, which is compounded by inefficient credit assignment in sequential prompt generation. To address this critical challenge, we develop a systematic approach in PAG with flow reactivation, reward-prioritized samplin g, and reward decomposition for prompt adaptation. Extensive experiments validat e that PAG successfully learns to sample effective and diverse prompts for text- to-image generation. We also show that PAG exhibits strong robustness across var ious reward functions and transferability to different text-to-image models.

****

Multi-modal Medical Diagnosis via Large-small Model Collaboration

Wanyi Chen, Zihua Zhao, Jiangchao Yao, Ya Zhang, Jiajun Bu, Haishuai Wang; Proce edings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, p p. 30763-30773

Recent advances in medical AI have shown a clear trend towards large models in h ealthcare. However, developing large models for multi-modal medical diagnosis re mains challenging due to a lack of sufficient modal-complete medical data. Most existing multi-modal diagnostic models are relatively small and struggle with li mited feature extraction capabilities. To bridge this gap, we propose **AdaCoMed **, an **ada**ptive **co**llaborative-learning framework that synergistically in tegrates the off-the-shelf **med**ical single-modal large models with multi-moda l small models. Our framework first employs a mixture-of-modality-experts (MoME) architecture to combine features extracted from multiple single-modal medical l arge models, and then introduces a novel adaptive co-learning mechanism to colla borate with a multi-modal small model. This co-learning mechanism, guided by an adaptive weighting strategy, dynamically balances the complementary strengths be tween the MoME-fused large model features and the cross-modal reasoning capabili ties of the small model. Extensive experiments on two representative multi-modal medical datasets (MIMIC-IV-MM and MMIST ccRCC) across six modalities and four d iagnostic tasks demonstrate consistent improvements over state-of-the-art baseli nes, making it a promising solution for real-world medical diagnosis application s.

****

SAMBLE: Shape-Specific Point Cloud Sampling for an Optimal Trade-Off Between Loc al Detail and Global Uniformity

Chengzhi Wu, Yuxin Wan, Hao Fu, Julius Pfrommer, Zeyun Zhong, Junwei Zheng, Jiam ing Zhang, Jürgen Beyerer; Proceedings of the Computer Vision and Pattern Recogn ition Conference (CVPR), 2025, pp. 1342-1352

Driven by the increasing demand for accurate and efficient representation of 3D data in various domains, point cloud sampling has emerged as a pivotal research topic in 3D computer vision. Recently, learning-to-sample methods have garnered growing interest from the community, particularly for their ability to be jointl y trained with downstream tasks. However, previous learning-based sampling metho ds either lead to unrecognizable sampling patterns by generating a new point clo ud or biased sampled results by focusing excessively on sharp edge details. More over, they all overlook the natural variations in point distribution across diff

erent shapes, applying a similar sampling strategy to all point clouds. In this paper, we propose a Sparse Attention Map and Bin-based Learning method (termed SAMBLE) to learn shape-specific sampling strategies for point cloud shapes. SAMBLE effectively achieves an improved balance between sampling edge points for local details and preserving uniformity in the global shape, resulting in superior performance across multiple common point cloud downstream tasks, even in scenarios with few-point sampling.

****

Image Referenced Sketch Colorization Based on Animation Creation Workflow

Dingkun Yan, Xinrui Wang, Zhuoru Li, Suguru Saito, Yusuke Iwasawa, Yutaka Matsuo, Jiaxian Guo; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23391-23400

Sketch colorization plays an important role in animation and digital illustration production tasks. However, existing methods still meet problems in that text-guided methods fail to provide accurate color and style reference, hint-guided methods still involve manual operation, and image-referenced methods are prone to cause artifacts. To address these limitations, we propose a diffusion-based framework inspired by real-world animation production workflows. Our approach leverages the sketch as the spatial guidance and an RGB image as the color reference, and separately extracts foreground and background from the reference image with spatial masks. Particularly, we introduce a split cross-attention mechanism with LoRA (Low-Rank Adaptation) modules. They are trained separately with foreground and background regions to control the corresponding embeddings for keys and values in cross-attention. This design allows the diffusion model to integrate information from foreground and background independently, preventing interference and eliminating the spatial artifacts. During inference, we design switchable inference modes for diverse use scenarios by changing modules activated in the framework. Extensive qualitative and quantitative experiments, along with user studies, demonstrate our advantages over existing methods in generating high-qualigy artifact-free results with geometric mismatched references. Ablation studies further confirm the effectiveness of each component. Codes are available at https://github.com/tellurion-kanata/colorizeDiffusion.

****

HoVLE: Unleashing the Power of Monolithic Vision-Language Models with Holistic Vision-Language Embedding

Chenxin Tao, Shiqian Su, Xizhou Zhu, Chenyu Zhang, Zhe Chen, Jiawen Liu, Wenhai Wang, Lewei Lu, Gao Huang, Yu Qiao, Jifeng Dai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14559-14569

The rapid advance of Large Language Models (LLMs) has catalyzed the development of Vision-Language Models (VLMs). Monolithic VLMs, which avoid modality-specific encoders, offer a promising alternative to the compositional ones but face the challenge of inferior performance. Most existing monolithic VLMs require tuning pre-trained LLMs to acquire vision abilities, which may degrade their language capabilities. To address this dilemma, this paper presents a novel high-performance monolithic VLM named HoVLE. We note that LLMs have been shown to be capable of interpreting images when image embeddings are aligned with text embeddings. The challenge for current monolithic VLMs actually lies in the lack of a holistic embedding module for both vision and language inputs. Therefore, HoVLE introduces a holistic embedding module that converts visual and textual inputs into a shared space, allowing LLMs to process images in the same way as texts. Furthermore, a multi-stage training strategy is carefully designed to empower the holistic embedding module. It is first trained to distill visual features from a pre-trained vision encoder and text embeddings from the LLM, enabling large-scale training with unpaired random images and text tokens. The whole model further undergoes next-token prediction on multi-modal data to align the embeddings. Finally, an instruction-tuning stage is incorporated. Our experiments show that HoVLE achieves performance close to leading compositional models on various benchmarks, outperforming previous monolithic models by a large margin.

****

Gen3DEval: Using vLLMs for Automatic Evaluation of Generated 3D Objects

Shalini Maiti, Lourdes Agapito, Filippos Kokkinos; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18552-18562

Rapid advancements in text-to-3D generation require robust and scalable evaluation metrics that align closely with human judgment, a need unmet by current metrics such as PSNR and CLIP, which require ground-truth data or focus only on prompt fidelity. To address this, we introduce Gen3DEval, a novel evaluation framework that leverages vision large language models (vLLMs) specifically fine-tuned for 3D object quality assessment. Gen3DEval evaluates text fidelity, appearance, and surface quality by analyzing 3D surface normals, without requiring ground-truth comparisons, bridging the gap between automated metrics and user preferences. Compared to state-of-the-art task-agnostic models, Gen3DEval demonstrates superior performance in user-aligned evaluations, placing it as a comprehensive and accessible benchmark for future research on text-to-3D generation. The project page can be found here: https://shalini-maiti.github.io/gen3deval.github.io/.

****

SkySense-O: Towards Open-World Remote Sensing Interpretation with Vision-Centric Visual-Language Modeling

Qi Zhu, Jiangwei Lao, Deyi Ji, Junwei Luo, Kang Wu, Yingying Zhang, Lixiang Ru, Jian Wang, Jingdong Chen, Ming Yang, Dong Liu, Feng Zhao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14733-14744

Open-world interpretation aims to accurately localize and recognize all objects within images by vision-language models (VLMs). While substantial progress has been made in this task for natural images, the advancements for remote sensing (RS) images still remain limited, primarily due to these two challenges. 1) Existing RS semantic categories are limited, particularly for pixel-level interpretation datasets. 2) Distinguishing among diverse RS spatial regions solely by language space is challenging due to the dense and intricate spatial distribution in open-world RS imagery. To address the first issue, we develop a fine-grained RS interpretation dataset, Sky-SA, which contains 183,375 high-quality local image-text pairs with full-pixel manual annotations, covering 1,763 category labels, exhibiting richer semantics and higher density than previous datasets. Afterwards, to solve the second issue, we introduce the vision-centric principle for vision-language modeling. Specifically, in the pre-training stage, the visual self-supervised paradigm is incorporated into image-text alignment, reducing the degradation of general visual representation capabilities of existing paradigms. Then, we construct a visual-relevance knowledge graph across open-category texts and further develop a novel vision-centric image-text contrastive loss for fine-tuning with text prompts. This new model, denoted as SkySense-O, demonstrates impressive zero-shot capabilities on a thorough evaluation encompassing 14 datasets over 4 tasks, from recognizing to reasoning and classification to localization. Specifically, it outperforms the latest models such as SegEarth-OV, GeoRSCLIP, and VHM by a large margin, i.e., 11.95%, 8.04% and 3.55% on average respectively.

****

AdaDARE-gamma: Balancing Stability and Plasticity in Multi-modal LLMs through Efficient Adaptation

Jingyi Xie, Jintao Yang, Zhunchen Luo, Yunbo Cao, Qiang Gao, Mengyuan Zhang, Wenpeng Hu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19758-19768

Adapting Multi-modal Large Language Models (MLLMs) to target tasks often suffers from catastrophic forgetting, where acquiring new task-specific knowledge compromises performance on pre-trained tasks. In this paper, we introduce AdaDARE-\gamma, an efficient approach that alleviates catastrophic forgetting by controllably injecting new task-specific knowledge through adaptive parameter selection from fine-tuned models without requiring retraining procedures. This approach consists two key innovations: (1) an adaptive parameter selection mechanism that identifies and retains the most task-relevant parameters from fine-tuned models, and (2) a controlled task-specific information injection strategy that precisely balances the preservation of pre-trained knowledge with the acquisition of new capabilities. Theoretical analysis proves the optimality of our parameter selection strategy and establishes bounds for the task-specific information injection fa

ctor. Extensive experiments on InstructBLIP and LLaVA-1.5 across image captionin g and visual question answering tasks demonstrate that AdaDARE-\gamma establishe s new state-of-the-art results in balancing model performance. Specifically, it maintains 98.2% of pre-training effectiveness on original tasks while achieving 98.7% of standard fine-tuning performance on target tasks.
****

Driving by the Rules: A Benchmark for Integrating Traffic Sign Regulations into Vectorized HD Map
Xinyuan Chang, Maixuan Xue, Xinran Liu, Zheng Pan, Xing Wei; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6823-6833
Ensuring adherence to traffic sign regulations is essential for both human and a utonomous vehicle navigation. While current online mapping solutions often prior itize the construction of the geometric and connectivity layers of HD maps, over looking the construction of the traffic regulation layer within HD maps. Address ing this gap, we introduce MapDR, a novel dataset designed for the extraction of Driving Rules from traffic signs and their association with vectorized, locally perceived HD Maps. MapDR features over 10,000 annotated video clips that captur e the intricate correlation between traffic sign regulations and lanes. Built up on this benchmark and the newly defined task of integrating traffic regulations into online HD maps, we provide modular and end-to-end solutions: VLE-MEE and Ru leVLM, offering a strong baseline for advancing autonomous driving technology. I t fills a critical gap in the integration of traffic sign rules, contributing to the development of reliable autonomous driving systems.
****

LeviTor: 3D Trajectory Oriented Image-to-Video Synthesis
Hanlin Wang, Hao Ouyang, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Qifeng Chen, Yuju n Shen, Limin Wang; Proceedings of the Computer Vision and Pattern Recognition C onference (CVPR), 2025, pp. 12490-12500
The intuitive nature of drag-based interaction has led to its growing adoption f or controlling object trajectories in image-to-video synthesis. Still, existing methods that perform dragging in the 2D space usually face ambiguity when handli ng out-of-plane movements. In this work, we augment the interaction with a new d imension, i.e., the depth dimension, such that users are allowed to assign a rel ative depth for each point on the trajectory. That way, our new interaction para digm not only inherits the convenience from 2D dragging, but facilitates traject ory control in the 3D space, broadening the scope of creativity. We propose a pi oneering method for 3D trajectory control in image-to-video synthesis by abstrac ting object masks into a few cluster points. These points, accompanied by the de pth information and the instance information, are finally fed into a video diffu sion model as the control signal. Extensive experiments validate the effectivene ss of our approach, dubbed LeviTor, in precisely manipulating the object movemen ts when producing photo-realistic videos from static images. Our code is availab le at: https://github.com/ant-research/LeviTor.
****

GaPT-DAR: Category-level Garments Pose Tracking via Integrated 2D Deformation an d 3D Reconstruction
Li Zhang, Mingliang Xu, Jianan Wang, Qiaojun Yu, Lixin Yang, Yonglu Li, Cewu Lu, Rujing Wang, Liu Liu; Proceedings of the Computer Vision and Pattern Recognitio n Conference (CVPR), 2025, pp. 22638-22647
Garments are common in daily life and are important for embodied intelligence co mmunity. Current category-level garments pose tracking works focus on predicting point-wise canonical correspondence and learning a shape deformation in point c loud sequences. In this paper, motivated by the 2D warping space and shape prior , we propose GaPT-DAR, a novel category-level Garments Pose Tracking framework w ith integrated 2D Deformation And 3D Reconstruction function, which fully utiliz e 3D-2D projection and 2D-3D reconstruction to transform the 3D point-wise learn ing into 2D warping deformation learning. Specifically, GaPT-DAR firstly builds a Voting-based Project module that learns the optimal 3D-2D projection plane for maintaining the maximum orthogonal entropy during point projection. Next, a Gar ments Deformation module is designed in 2D space to explicitly model the garment

s warping procedure with deformation parameters. Finally, we build a Depth Recon
struction module to recover the 2D images into 3D warp field. We provide extensi
ve experiments on VR-Folding dataset to evaluate our GaPT-DAR and the results sh
ow obvious improvements on most of the metrics compared to state-of-the-arts (G
armentNets and GarmentTracking). More details are available at https://sites.goo
gle.com/view/gapt-dar.
****

ProbPose: A Probabilistic Approach to 2D Human Pose Estimation
Miroslav Purkrabek, Jiri Matas; Proceedings of the Computer Vision and Pattern R
ecognition Conference (CVPR), 2025, pp. 27124-27133
Current state-of-the-art Human Pose Estimation methods ignore out-of-image keypo
ints in both training and evaluation and use uncalibrated heatmaps as keypoint l
ocation representations. We propose ProbPose, which predicts for each keypoint:
a calibrated probability of keypoint presence at each location in the activation
 window, the probability of being outside of it, and its predicted visibility. T
o address the lack of evaluation protocols for out-of-image keypoints, we introd
uce the CropCOCO dataset and the Extended OKS (Ex-OKS) metric, which extends OKS
 to out-of-image points. Tested on COCO, CropCOCO, and OCHuman, ProbPose shows s
ignificant gains in out-of-image keypoint localization while also improving in-i
mage localization through data augmentation. Additionally, the model improves ro
bustness along the edges of the bounding box and offers better flexibility in ke
ypoint evaluation. The code and weights are available on the MiraPurkrabek.githu
b.io/ProbPose.
****

SapiensID: Foundation for Human Recognition
Minchul Kim, Dingqiang Ye, Yiyang Su, Feng Liu, Xiaoming Liu; Proceedings of the
 Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13937-1394
7
Existing human recognition systems often rely on separate, specialized models fo
r face and body analysis, limiting their effectiveness in real-world scenarios w
here pose, visibility, and context vary widely. This paper introduces SapiensID,
 a unified model that bridges this gap, achieving robust performance across dive
rse settings. SapiensID introduces (i) Retina Patch (RP), a dynamic patch genera
tion scheme that adapts to subject scale and ensures consistent tokenization of
regions of interest; (ii) Semantic Attention Head (SAH), an attention mechanism
that learns pose-invariant representations by pooling features around key body p
arts; and (iii) a masked recognition model (MRM) that learns from variable token
 length. To facilitate training, we introduce WebBody4M, a large-scale dataset c
apturing diverse poses and scale variations. Extensive experiments demonstrate t
hat SapiensID achieves state-of-the-art results on various body ReID benchmarks,
 outperforming specialized models in both short-term and long-term scenarios whi
le remaining competitive with dedicated face recognition systems. Furthermore, S
apiensID establishes a strong baseline for the newly introduced challenge of Cro
ss Pose-Scale ReID, demonstrating its ability to generalize to complex, real-wor
ld conditions.The dataset, code and models will be released.
****

MIDI: Multi-Instance Diffusion for Single Image to 3D Scene Generation
Zehuan Huang, Yuan-Chen Guo, Xingqiao An, Yunhan Yang, Yangguang Li, Zi-Xin Zou,
 Ding Liang, Xihui Liu, Yan-Pei Cao, Lu Sheng; Proceedings of the Computer Visio
n and Pattern Recognition Conference (CVPR), 2025, pp. 23646-23657
This paper introduces MIDI, a novel paradigm for compositional 3D scene generati
on from a single image. Unlike existing methods that rely on reconstruction or r
etrieval techniques or recent approaches that employ multi-stage object-by-objec
t generation, MIDI extends pre-trained image-to-3D object generation models to m
ulti-instance diffusion models, enabling the simultaneous generation of multiple
 3D instances with accurate spatial relationships and high generalizability. At
its core, MIDI incorporates a novel multi-instance attention mechanism, that eff
ectively captures inter-object interactions and spatial coherence directly withi
n the generation process, without the need for complex multi-step processes. The
 method utilizes partial object images and global scene context as inputs, direc

tly modeling object completion during 3D generation. During training, we effecti vely supervise the interactions between 3D instances using a limited amount of s cene-level data, while incorporating single-object data for regularization, ther eby maintaining the pre-trained generalization ability. MIDI demonstrates state- of-the-art performance in image-to-scene generation, validated through evaluatio ns on synthetic data, real-world scene data, and stylized scene images generated by text-to-image diffusion models.
****
S4-Driver: Scalable Self-Supervised Driving Multimodal Large Language Model with Spatio-Temporal Visual Representation
Yichen Xie, Runsheng Xu, Tong He, Jyh-Jing Hwang, Katie Luo, Jingwei Ji, Hubert Lin, Letian Chen, Yiren Lu, Zhaoqi Leng, Dragomir Anguelov, Mingxing Tan; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1622-1632
The latest advancements in multi-modal large language models (MLLMs) have spurred a strong renewed interest in end-to-end motion planning approaches for autonomous driving. Many end-to-end approaches rely on human annotations to learn intermediate perception and prediction tasks, while purely self-supervised approaches --which directly learn from sensor inputs to generate planning trajectories without human annotations--often underperform the state of the art. We observe a key gap in the input representation space: end-to-end approaches built on MLLMs are often pretrained with reasoning tasks in 2D image space rather than the native 3D space in which autonomous vehicles plan. To this end, we propose S4-Driver, a scalable self-supervised motion planning algorithm with spatio-temporal visual representation, based on the popular PaLI multimodal large language model. S4-Driver uses a novel sparse volume strategy to seamlessly transform the strong visual representation of MLLMs from perspective view to 3D space without the need to finetune the vision encoder. This representation aggregates multi-view and multi-frame visual inputs and enables better prediction of planning trajectories in 3D space. To validate our method, we run experiments on both nuScenes and Waymo Open Motion Dataset (with in-house camera data). Results show that S4-Driver performs favorably against existing supervised multi-task approaches while requiring no human annotations. It also demonstrates great scalability when pretrained on large volumes of unannotated driving logs.
****
Extrapolating and Decoupling Image-to-Video Generation Models: Motion Modeling is Easier Than You Think
Jie Tian, Xiaoye Qu, Zhenyi Lu, Wei Wei, Sichen Liu, Yu Cheng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12512-12521
Image-to-Video (I2V) generation aims to synthesize a video clip according to a given image and condition (e.g., text). The key challenge of this task lies in simultaneously generating natural motions while preserving the original appearance of the images. However, current I2V diffusion models (I2V-DMs) often produce videos with limited motion degrees or exhibit uncontrollable motion that conflicts with the textual condition. To address these limitations, we propose a novel Extrapolating and Decoupling framework, which introduces model merging techniques to the I2V domain for the first time. Specifically, our framework consists of three separate stages: (1) Starting with a base I2V-DM, we explicitly inject the textual condition into the temporal module using a lightweight, learnable adapter and fine-tune the integrated model to improve motion controllability. (2) We introduce a training-free extrapolation strategy to amplify the dynamic range of the motion, effectively reversing the fine-tuning process to enhance the motion degree significantly. (3) With the above two-stage models excelling in motion controllability and degree, we decouple the relevant parameters associated with each type of motion ability and inject them into the base I2V-DM. Since the I2V-DM handles different levels of motion controllability and dynamics at various denoising time steps, we adjust the motion-aware parameters accordingly over time. Extensive qualitative and quantitative experiments have been conducted to demonstrate the superiority of our framework over existing methods.

****
FreeCloth: Free-form Generation Enhances Challenging Clothed Human Modeling
Hang Ye, Xiaoxuan Ma, Hai Ci, Wentao Zhu, Yizhou Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15987-15997
Achieving realistic animated human avatars requires accurate modeling of pose-dependent clothing deformations. Existing learning-based methods heavily rely on the Linear Blend Skinning (LBS) of minimally-clothed human models like SMPL to model deformation. However, they struggle to handle loose clothing, such as long dresses, where the canonicalization process becomes ill-defined when the clothing is far from the body, leading to disjointed and fragmented results. To overcome this limitation, we propose FreeCloth, a novel hybrid framework to model challenging clothed humans. Our core idea is to use dedicated strategies to model different regions, depending on whether they are close to or distant from the body. Specifically, we segment the human body into three categories: unclothed, deformed, and generated. We simply replicate unclothed regions that require no deformation. For deformed regions close to the body, we leverage LBS to handle the deformation. As for the generated regions, which correspond to loose clothing areas, we introduce a novel free-form, part-aware generator to model them, as they are less affected by movements. This free-form generation paradigm brings enhanced flexibility and expressiveness to our hybrid framework, enabling it to capture the intricate geometric details of challenging loose clothing, such as skirts and dresses. Experimental results on the benchmark dataset featuring loose clothing demonstrate that FreeCloth achieves state-of-the-art performance with superior visual fidelity and realism, particularly in the most challenging cases.
****
ABC-Former: Auxiliary Bimodal Cross-domain Transformer with Interactive Channel Attention for White Balance
Yu-Cheng Chiu, Guan-Rong Chen, Zihao Chen, Yan-Tsung Peng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 21258-21266
The primary goal of white balance (WB) for sRGB images is to correct inaccurate color temperatures, ensuring that images display natural, neutral colors. While existing WB methods yield reasonable results, their effectiveness is limited. They either focus solely on global color adjustments applied before the camera-specific image signal processing pipeline or rely on end-to-end models that generate WB outputs without accounting for global color trends, leading to suboptimal correction. To address these limitations, we propose an Auxiliary Bimodal Cross-domain Transformer (ABC-Former) that enhances WB correction by leveraging complementary knowledge from global color information from CIELab and RGB histograms alongside sRGB inputs. By introducing an Interactive Channel Attention (ICA) module to facilitate cross-modality global knowledge transfer, ABC-Former achieves more precise WB correction. Experimental results on benchmark WB datasets show that ABC-Former performs favorably against state-of-the-art WB methods. The source code is available at https://github.com/ytpeng-aimlab/ABC-Former.
****
Science-T2I: Addressing Scientific Illusions in Image Synthesis
Jialuo Li, Wenhao Chai, Xingyu Fu, Haiyang Xu, Saining Xie; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2734-2744
We present a novel approach to integrating scientific knowledge into generative models, enhancing their realism and consistency in image synthesis. First, we introduce Science-T2I, an expert-annotated adversarial dataset comprising adversarial 20k image pairs with 9k prompts, covering wide distinct scientific knowledge categories. Leveraging Science-T2I, we present SciScore, an end-to-end reward model that refines the assessment of generated images based on scientific knowledge, which is achieved by augmenting both the scientific comprehension and visual capabilities of pre-trained CLIP model. Additionally, based on Science-T2I, we propose a two-stage training framework, comprising a supervised fine-tuning phase and a masked online fine-tuning phase, to incorporate scientific knowledge into existing generative models. Through comprehensive experiments, we demonstrate the effectiveness of our framework in establishing new standards for evaluating the scientific realism of generated content. Specifically, SciScore attains perf

ormance comparable to human-level, demonstrating a 5% improvement similar to eva luations conducted by experienced human evaluators. Furthermore, by applying our proposed fine-tuning method to FLUX, we achieve a performance enhancement excee ding 50% on SciScore.

****

Fingerprinting Denoising Diffusion Probabilistic Models

Huan Teng, Yuhui Quan, Chengyu Wang, Jun Huang, Hui Ji; Proceedings of the Compu ter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 28811-28820

Diffusion models, especially denoising diffusion probabilistic models (DDPMs), a re prevalent tools in generative AI, making their intellectual property (IP) pro tection increasingly important. Most existing IP protection methods for DDPMs ar e invasive, e.g., model watermarking, which alter model parameters and raise con cerns about performance degradation, also with requirement for extra computation al resources for retraining or fine-tuning. In this paper, we propose the first non-invasive fingerprinting scheme for DDPMs, requiring no parameter changes or fine-tuning, and keeping generation quality intact. We introduce a discriminativ e and robust fingerprint latent space based on the well-designed "crossing route " of noisy samples that span the performance border-zone of DDPMs, with only bla ck-box access required for the diffusion denoiser in ownership verification. Ext ensive experiments demonstrate that our fingerprinting approach enjoys both robu stness against the often-seen attacks and distinctiveness on various DDPMs, prov iding an alternative for protecting DDPMs' IP rights without compromising their performance or integrity.

****

MoST: Efficient Monarch Sparse Tuning for 3D Representation Learning

Xu Han, Yuan Tang, Jinfeng Xu, Xianzhi Li; Proceedings of the Computer Vision an d Pattern Recognition Conference (CVPR), 2025, pp. 6584-6594

We introduce Monarch Sparse Tuning (MoST), the first reparameterization-based pa rameter-efficient fine-tuning (PEFT) method tailored for 3D representation learn ing. Unlike existing adapter-based and prompt-tuning 3D PEFT methods, MoST intro duces no additional inference overhead and is compatible with many 3D representa tion learning backbones. At its core, we present a new family of structured matr ices for 3D point clouds, Point Monarch, which can capture local geometric featu res of irregular points while offering high expressiveness. MoST reparameterizes the dense update weight matrices as our sparse Point Monarch matrices, signific antly reducing parameters while retaining strong performance. Experiments on var ious backbones show that MoST is simple, effective, and highly generalizable. It captures local features in point clouds, achieving state-of-the-art results on multiple benchmarks, e.g., 97.5% acc. on ScanObjectNN (PB_50_RS) and 96.2% on Mo delNet40 classification, while it can also combine with other matrix decompositi ons (e.g., Low-rank, Kronecker) to further reduce parameters.

****

Re-thinking Temporal Search for Long-Form Video Understanding

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cris tobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, Ji ajun Wu, Manling Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8579-8591

Efficient understanding of long-form videos remains a significant challenge in c omputer vision. In this work, we revisit temporal search paradigms for long-form video understanding, studying a fundamental issue pertaining to all state-of-th e-art (SOTA) long-context vision-language models (VLMs). In particular, our cont ributions are two-fold: **First**, we formulate temporal search as a **Long Vide o Haystack** problem, i.e., finding a minimal set of relevant frames (typically one to five) among tens of thousands of frames from real-world long videos given specific queries. To validate our formulation, we create **LV-Haystack**, the f irst benchmark containing 3,874 human-annotated instances with fine-grained eval uation metrics for assessing keyframe search quality and computational efficienc y. Experimental results on LV-Haystack highlight a significant research gap in t emporal search capabilities, with SOTA keyframe selection methods achieving only 2.1% temporal F1 score on the LVBench subset.**Next**, inspired by visual searc

h in images, we re-think temporal searching and propose a lightweight keyframe s earching framework, T^* , which casts the expensive temporal search as a spatial search problem. T^* leverages superior visual localization capabilities typica lly used in images and introduces an adaptive zooming-in mechanism that operates across both temporal and spatial dimensions. Our extensive experiments show tha t when integrated with existing methods, T^* significantly improves SOTA long-f orm video understanding performance. Specifically, under an inference budget of 32 frames, T^* improves GPT-4o's performance from 50.5% to **53.1%** and LLaVA-OneVision-72B's performance from 56.5% to **62.4%** on LongVideoBench XL subset. Our PyTorch code, benchmark dataset and models are included in the Supplementar y material.

****

InstanceGaussian: Appearance-Semantic Joint Gaussian Representation for 3D Insta nce-Level Perception

Haijie Li, Yanmin Wu, Jiarui Meng, Qiankun Gao, Zhiyao Zhang, Ronggang Wang, Jia n Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 14078-14088

3D scene understanding is vital for applications in autonomous driving, robotics , and augmented reality. However, scene understanding based on 3D Gaussian Splat ting faces three key challenges: (i) an imbalance between appearance and semanti cs, (ii) inconsistencies in object boundaries, and (iii) difficulties with top-d own instance segmentation. To address these challenges, we propose InstanceGauss ian, a method that jointly learns appearance and semantic features while adaptiv ely aggregating instances. Our contributions are as follows: (i) a new Semantic-Scaffold-GS representation to improve feature representation and boundary deline ation, (ii) a progressive training strategy for enhanced stability and segmentat ion, and (iii) a category-agnostic, bottom-up instance aggregation approach for better segmentation. Experimental results demonstrate that our approach achieves state-of-the-art performance in category-agnostic, open-vocabulary 3D point-lev el segmentation, validating the effectiveness of our proposed method.

****

When Domain Generalization meets Generalized Category Discovery: An Adaptive Tas k-Arithmetic Driven Approach

Vaibhav Rathore, Shubhranil B, Saikat Dutta, Sarthak Mehrotra, Zsolt Kira, Bipla b Banerjee; Proceedings of the Computer Vision and Pattern Recognition Conferenc e (CVPR), 2025, pp. 4905-4915

Generalized Class Discovery (GCD) clusters base and novel classes in a target do main, using supervision from a source domain with only base classes. Current met hods often falter with distribution shifts and typically require access to targe t data during training, which can sometimes be impractical. To address this issu e, we introduce the novel paradigm of Domain Generalization in GCD (DG-GCD), whe re only source data is available for training, while the target domain--with a d istinct data distribution--remains unseen until inference. To this end, our solu tion, DG2CD-Net, aims to construct a domain-independent, discriminative embeddin g space for GCD. The core innovation is an episodic training strategy that enhan ces cross-domain generalization by adapting a base model on tasks derived from s ource and synthetic domains generated by a foundation model. Each episode focuse s on a cross-domain GCD task, diversifying task setups over episodes and combini ng open-set domain adaptation with a novel margin loss and representation learni ng for optimizing the feature space progressively. To capture the effects of fin e-tunings on the base model, we extend task arithmetic by adaptively weighting t he local task vectors concerning the fine-tuned models based on their GCD perfor mance on a validation distribution. This episodic update mechanism boosts the ad aptability of the base model to unseen targets. Experiments across three dataset s confirm that DG2CD-Net outperforms existing GCD methods customized for DG-GCD

****

CSC-PA: Cross-image Semantic Correlation via Prototype Attentions for Single-net work Semi-supervised Breast Tumor Segmentation

Zhenhui Ding, Guilian Chen, Qin Zhang, Huisi Wu, Jing Qin; Proceedings of the Co mputer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15632-15641

Accurate automatic breast ultrasound (BUS) image segmentation is essential for e
arly breast cancer screening and diagnosis. However, it remains challenging owin
g to (1) breast lesions of various scale and shape, (2) ambiguous boundaries cau
sed by speckle noise and artifacts, and (3) the scarcity of high-quality annotat
ions. Most existing semi-supervised methods employ the mean-teacher architecture
, which merely learns semantic information within a single image and heavily rel
ies on the performance of the teacher model. Therefore, we present a novel cross
-image semantic correlation semi-supervised framework, named CSC-PA, to improve
the performance of BUS image segmentation. CSC-PA is trained based on a single n
etwork, which integrates a foreground prototype attention (FPA) and an edge prot
otype attention (EPA). Specifically, FPA transfers complementary foreground info
rmation for more stable and complete lesion segmentation. On the other hand, EPA
 enhances edge features of lesions by using edge prototype, where an adaptive ed
ge container is proposed to store global edge features and generate the edge pro
totype. Additionally, we introduce a pixel affinity loss (PAL) to exploit previo
usly ignored contextual correlation in supervision, which further improves perfo
rmance on edges. Extensive experiments on two benchmark BUS datasets demonstrate
 that our model outperforms other state-of-the-art methods under different parti
tion protocols. Codes are available at https://github.com/shdkdh/CSC-PA.
****

BIP3D: Bridging 2D Images and 3D Perception for Embodied Intelligence
Xuewu Lin, Tianwei Lin, Lichao Huang, Hongyu Xie, Zhizhong Su; Proceedings of th
e Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9007-9016
In embodied intelligence systems, a key component is 3D perception algorithm, wh
ich enables agents to understand their surrounding environments. Previous algori
thms primarily rely on point cloud, which, despite offering precise geometric in
formation, still constrain perception performance due to inherent sparsity, nois
e, and data scarcity. In this work, we introduce a novel image-centric 3D percep
tion model, BIP3D, which leverages expressive image features with explicit 3D po
sition encoding to overcome the limitations of point-centric methods. Specifical
ly, we leverage pre-trained 2D vision foundation models to enhance semantic unde
rstanding, and introduce a spatial enhancer module to improve spatial understand
ing. Together, these modules enable BIP3D to achieve multi-view, multi-modal fea
ture fusion and end-to-end 3D perception. In our experiments, BIP3D outperforms
current state-of-the-art results on the EmbodiedScan benchmark, achieving improv
ements of 5.69% in the 3D detection task and 15.25% in the 3D visual grounding t
ask. Code has been released at https://github.com/HorizonRobotics/BIP3D.
****

Query Efficient Black-Box Visual Prompting with Subspace Learning
Zhaogeng Liu, Haozhen Zhang, Hualin Zhang, Xingchen Li, Wanli Shi, Bin Gu, Yi Ch
ang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR
), 2025, pp. 4322-4331
Visual Prompt Learning (VPL) has emerged as a powerful strategy for harnessing t
he capabilities of large-scale pre-trained models (PTMs) to tackle specific down
stream tasks. However, the opaque nature of PTMs in many real-world applications
 has led to a growing interest in gradient-free approaches within VPL. A signifi
cant challenge with existing black-box VPL methods lies in the high dimensionali
ty of visual prompts, which necessitates considerable API queries for tuning, th
ereby impacting efficiency. To address this issue, we propose a novel query-effi
cient framework for black-box visual prompting, designed to generate input-depen
dent visual prompts efficiently for large-scale black-box PTMs. Our framework is
 built upon the insight of reparameterizing prompts using neural networks, impro
ving the typical pre-training-fine-tuning paradigm through the subspace learning
 strategy to maximize efficiency and adaptability from both the perspective of i
nitial weights and parameter dimensionality. This tuning intrinsically optimizes
 low-dimensional representations within the well-learned subspace, enabling the
efficient adaptation of the network to downstream tasks. Our approach significan
tly reduces the necessity for substantial API queries to PTMs, presenting an eff
icient method for leveraging large-scale black-box PTMs in visual prompting task
s. Most experimental results across various benchmarks demonstrate the effective

ness of our method, showcasing substantial reductions in the number of required API queries to PTMs while maintaining or even enhancing performance on downstrea m tasks.
****

VisionPAD: A Vision-Centric Pre-training Paradigm for Autonomous Driving
Haiming Zhang, Wending Zhou, Yiyao Zhu, Xu Yan, Jiantao Gao, Dongfeng Bai, Yingj ie Cai, Bingbing Liu, Shuguang Cui, Zhen Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 17165-17175
This paper introduces VisionPAD, a novel self-supervised pre-training paradigm d esigned for vision-centric algorithms in autonomous driving. In contrast to prev ious approaches that employ neural rendering with explicit depth supervision, Vi sionPAD utilizes more efficient 3D Gaussian Splatting to reconstruct multi-view representations using only images as supervision. Specifically, we introduce a s elf-supervised method for voxel velocity estimation. By warping voxels to adjace nt frames and supervising the rendered outputs, the model effectively learns mot ion cues in the sequential data. Furthermore, we adopt a multi-frame photometric consistency approach to enhance geometric perception. It projects adjacent fram es to the current frame based on rendered depths and relative poses, boosting th e 3D geometric representation through pure image supervision. Extensive experime nts on autonomous driving datasets demonstrate that VisionPAD significantly impr oves performance in 3D object detection, occupancy prediction and map segmentati on, surpassing state-of-the-art pre-training strategies by a considerable margin .
****

Detecting Adversarial Data Using Perturbation Forgery
Qian Wang, Chen Li, Yuchen Luo, Hefei Ling, Shijuan Huang, Ruoxi Jia, Ning Yu; P roceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 202 5, pp. 13917-13926
As a defense strategy against adversarial attacks, adversarial detection aims to identify and filter out adversarial data from the data flow based on discrepanc ies in distribution and noise patterns between natural and adversarial data. Alt hough previous detection methods achieve high performance in detecting gradient-based adversarial attacks, new attacks based on generative models with imbalance d and anisotropic noise patterns evade detection. Even worse, the significant in ference time overhead and limited performance against unseen attacks make existi ng techniques impractical for real-world use. In this paper, we explore the prox imity relationship among adversarial noise distributions and demonstrate the exi stence of an open covering for these distributions. By training on the open cove ring of adversarial noise distributions, a detector with strong generalization p erformance against various types of unseen attacks can be developed. Based on th is insight, we heuristically propose Perturbation Forgery, which includes noise distribution perturbation, sparse mask generation, and pseudo-adversarial data p roduction, to train an adversarial detector capable of detecting any unseen grad ient-based, generative-based, and physical adversarial attacks. Comprehensive ex periments conducted on multiple general and facial datasets, with a wide spectru m of attacks, validate the strong generalization of our method.
****

CoA: Towards Real Image Dehazing via Compression-and-Adaptation
Long Ma, Yuxin Feng, Yan Zhang, Jinyuan Liu, Weimin Wang, Guang-Yong Chen, Cheng pei Xu, Zhuo Su; Proceedings of the Computer Vision and Pattern Recognition Conf erence (CVPR), 2025, pp. 11197-11206
Learning-based image dehazing algorithms have shown remarkable success in synthe tic domains. However, real image dehazing is still in suspense due to computatio nal resource constraints and the diversity of real-world scenes. Therefore, ther e is an urgent need for an algorithm that excels in both efficiency and adaptabi lity to address real image dehazing effectively. This work proposes a Compressio n-and-Adaptation (CoA) computational flow to tackle these challenges from a divi de-and-conquer perspective. First, model compression is performed in the synthet ic domain to develop a compact dehazing parameter space, satisfying efficiency d emands. Then, a bilevel adaptation in the real domain is introduced to be fearle

ss in unknown real environments by aggregating the synthetic dehazing capabiliti es during the learning process. Leveraging a succinct design free from additiona l constraints, our CoA exhibits domain-irrelevant stability and model-agnostic f lexibility, effectively bridging the model chasm between synthetic and real dom ains to further improve its practical utility. Extensive evaluations and analyse s underscore the approach's superiority and effectiveness. The code is publicly available at https://github.com/fyxnl/COA.
****

NightAdapter: Learning a Frequency Adapter for Generalizable Night-time Scene Se gmentation
Qi Bi, Jingjun Yi, Huimin Huang, Hao Zheng, Haolan Zhan, Yawen Huang, Yuexiang L i, Xian Wu, Yefeng Zheng; Proceedings of the Computer Vision and Pattern Recogni tion Conference (CVPR), 2025, pp. 23838-23849
Night-time scene segmentation is a critical yet challenging task in the real-wor ld applications, primarily due to the complicated lighting conditions. However, existing methods lack sufficient generalization ability to unseen nigh-time scen es with varying illumination.In light of this issue, we focus on investigating g eneralizable paradigms for night-time scene segmentation and propose an efficien t fine-tuning scheme, dubbed as \texttt NightAdapter , alleviating the domain ga p across various scenes.Interestingly, different properties embedded in the day- time and night-time features can be characterized by the bands after discrete si ne transformation, which can be categorized into illumination-sensitive/-insensi tive bands.Hence, our \texttt NightAdapter is powered by two appealing designs: (1) Illumination-Insensitive Band Adaptation that provides a foundation for und erstanding the prior, enhancing the robustness to illumination shifts; (2) Illum ination-Sensitive Band Adaptation that fine-tunes the randomized frequency bands , mitigating the domain gap between the day-time and various night-time scenes. As a consequence, illumination-insensitive enhancement improves the domain invar iance, while illumination-sensitive diminution strengthens the domain shift betw een different scenes.\texttt NightAdapter yields significant improvements over the state-of-the-art methods under various day-to-night, night-to-night, and in- domain night segmentation experiments.We will release our code.
****

UMFN: Unified Multi-Domain Face Normalization for Joint Cross-domain Prototype L earning and Heterogeneous Face Recognition
Meng Pang, Wenjun Zhang, Nanrun Zhou, Shengbo Chen, Hong Rao; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29299-2930 8
Face normalization aims to enhance the robustness and effectiveness of face reco gnition systems by mitigating intra-personal variations in expressions, poses, o cclusions, illuminations, and domains. Existing methods face limitations in hand ling multiple variations and adapting to cross-domain scenarios. To address thes e challenges, we propose a novel Unified Multi-Domain Face Normalization Network (UMFN) model, which can process face images with various types of facial variat ions from different domains, and reconstruct frontal, neutral-expression facial prototypes in the target domain. As an unsupervised domain adaptation model, UMF N facilitates concurrent training on multiple datasets across domains and demons trates strong prototype reconstruction capabilities. Notably, UMFN serves as a j oint prototype and feature learning framework, enabling the simultaneous extract ion of domain-agnostic identity features through a decoupling mapping network an d a feature domain classifier for adversarial training. Moreover, we design an e fficient Heterogeneous Face Recognition (HFR) network that fuses domain-agnostic and identity-discriminative features for HFR, and introduce contrastive learnin g to enhance identity recognition accuracy. Empirical studies on diverse cross-d omain face datasets validate the effectiveness of our proposed method.
****

TopV: Compatible Token Pruning with Inference Time Optimization for Fast and Low -Memory Multimodal Vision Language Model
Cheng Yang, Yang Sui, Jinqi Xiao, Lingyi Huang, Yu Gong, Chendi Li, Jinghua Yan, Yu Bai, Ponnuswamy Sadayappan, Xia Hu, Bo Yuan; Proceedings of the Computer Vis

ion and Pattern Recognition Conference (CVPR), 2025, pp. 19803-19813
Vision-Language Models (VLMs) demand substantial computational resources during inference, largely due to the extensive visual input tokens for representing visual information. Previous studies have noted that visual tokens tend to receive less attention than text tokens, suggesting their lower importance during inference and potential for pruning. However, their methods encounter several chal- lenges: reliance on greedy heuristic criteria for token im- portance and incompatibility with FlashAttention and KV cache. To address these issues, we introduce TopV, a com- patible TOken Pruning with inference Time Optimization for fast and low-memory VLM, achieving efficient pruning without additional training or fine-tuning. Instead of re- lying on attention scores, we formulate token pruning as an optimization problem, accurately identifying important visual tokens while remaining compatible with FlashAtten- tion. Additionally, since we only perform this pruning once during the prefilling stage, it effectively reduces KV cache size. Our optimization framework incorporates a visual- aware cost function considering factors such as Feature Similarity, Relative Spatial Distance, and Absolute Central Distance, to measure the importance of each source visual token, enabling effective pruning of low-importance tokens. Extensive experiments demonstrate that our method outperforms previous token pruning methods, validating the effectiveness and efficiency of our approach.
****

Improving Autoregressive Visual Generation with Cluster-Oriented Token Prediction

Teng Hu, Jiangning Zhang, Ran Yi, Jieyu Weng, Yabiao Wang, Xianfang Zeng, Zhucun Xue, Lizhuang Ma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9351-9360
Employing LLMs for visual generation has recently become a research focus. However, the existing methods primarily transfer the LLM architecture to visual generation but rarely investigate the fundamental differences between language and vision. This oversight may lead to suboptimal utilization of visual generation capabilities within the LLM framework. In this paper, we explore the characteristics of visual embedding space under the LLM framework and discover that the correlation between visual embeddings can help achieve more stable and robust generation results. We present **IAR**, an **I**mproved **A**uto**R**egressive Visual Generation Method that enhances the training efficiency and generation quality of LLM-based visual generation models. Firstly, we propose a Codebook Rearrangement strategy that uses balanced k-means clustering algorithm to rearrange the visual codebook into clusters, ensuring high similarity among visual features within each cluster. Leveraging the rearranged codebook, we propose a Cluster-oriented Cross-entropy Loss that guides the model to correctly predict the cluster where the token is located. This approach ensures that even if the model predicts the wrong token index, there is a high probability the predicted token is located in the correct cluster, which significantly enhances the generation quality and robustness. Extensive experiments demonstrate that our method consistently enhances the model training efficiency and performance from 100M to 1.4B, reducing the training time by half while achieving the same FID. Additionally, our approach can be applied to various LLM-based visual generation models and adheres to the scaling law, providing a promising direction for future research in LLM-based visual generation.
****

Learned Binocular-Encoding Optics for RGBD Imaging Using Joint Stereo and Focus Cues

Yuhui Liu, Liangxun Ou, Qiang Fu, Hadi Amata, Wolfgang Heidrich, Yifan Peng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15833-15842
Extracting high-fidelity RGBD information from two-dimensional (2D) images is essential for various visual computing applications. Stereo imaging, as a reliable passive imaging technique for obtaining three-dimensional (3D) scene information, has benefited greatly from deep learning advancements. However, existing stereo depth estimation algorithms struggle to perceive high-frequency information a

nd resolve high-resolution depth maps in realistic camera settings with large de
pth variations. These algorithms commonly neglect the hardware parameter configu
ration, limiting the potential for achieving optimal solutions solely through so
ftware-based design strategies.This work presents a hardware-software co-designe
d RGBD imaging framework that leverages both stereo and focus cues to reconstruc
t texture-rich color images along with detailed depth maps over a wide depth ran
ge. A pair of rank-2 parameterized diffractive optical elements (DOEs) is employ
ed to encode perpendicular complementary information optically during stereo acq
uisitions. Additionally, we employ an IGEV-UNet-fused neural network tailored to
 the proposed rank-2 encoding for stereo matching and image reconstruction. Thro
ugh prototyping a stereo camera with customized DOEs, our deep stereo imaging pa
radigm has demonstrated superior performance over existing monocular and stereo
 imaging systems in both image PSNR by 2.96 dB gain and depth accuracy in high-fr
equency details across distances from 0.67 to 8 meters.
****

Dual-view X-ray Detection: Can AI Detect Prohibited Items from Dual-view X-ray I
mages like Humans?
Renshuai Tao, Haoyu Wang, Yuzhe Guo, Hairong Chen, Li Zhang, Xianglong Liu, Yunc
hao Wei, Yao Zhao; Proceedings of the Computer Vision and Pattern Recognition Co
nference (CVPR), 2025, pp. 10338-10347
To detect prohibited items in challenging categories, human inspectors typically
 rely on images from two distinct views (vertical and side). Can AI detect prohi
bited items from dual-view X-ray images in the same way humans do? Existing X-ra
y datasets often suffer from limitations, such as single-view imaging or insuffi
cient sample diversity. To address these gaps, we introduce the Large-scale Dual
-view X-ray (LDXray), which consists of 353,646 instances across 12 categories,
providing a diverse and comprehensive resource for training and evaluating model
s. To emulate human intelligence in dual-view detection, we propose the Auxiliar
y-view Enhanced Network (AENet), a novel detection framework that leverages both
 the main and auxiliary views of the same object. The main-view pipeline focuses
 on detecting common categories, while the auxiliary-view pipeline handles more
challenging categories using "expert models" learned from the main view. Extensi
ve experiments on the LDXray dataset demonstrate that the dual-view mechanism si
gnificantly enhances detection performance, e.g., achieving improvements of up t
o +24.7%for the challenging category of umbrellas. Furthermore, our results show
 that AENet exhibits strong generalization across seven different detection mode
ls.
****

LUCAS: Layered Universal Codec Avatars
Di Liu, Teng Deng, Giljoo Nam, Yu Rong, Stanislav Pidhorskyi, Junxuan Li, Jason
Saragih, Dimitris N. Metaxas, Chen Cao; Proceedings of the Computer Vision and P
attern Recognition Conference (CVPR), 2025, pp. 21127-21137
Photorealistic 3D head avatar reconstruction faces critical challenges in modeli
ng dynamic face-hair interactions and achieving cross-identity generalization, p
articularly during expressions and head movements. We present LUCAS, a novel Uni
versal Prior Model (UPM) for codec avatar modeling that disentangles face and ha
ir through a layered representation. Unlike previous UPMs that treat hair as an
integral part of the head, our approach separates the modeling of the hairless h
ead and hair into distinct branches. LUCAS is the first to introduce a mesh-base
d UPM, facilitating real-time rendering on devices. LUCAS can be integrated with
 Gaussian Splatting to enhance visual fidelity, a feature particularly beneficia
l for rendering complex hairstyles. Experimental results indicate that LUCAS out
performs existing single-mesh and Gaussian-based avatar models in both quantitat
ive and qualitative assessments, including evaluations on held-out subjects in z
ero-shot driving scenarios. LUCAS demonstrates superior dynamic performance in m
anaging head pose changes, expression transfer, and hairstyle variations, thereb
y advancing the state-of-the-art in 3D head avatar reconstruction.
****

MobilePortrait: Real-Time One-Shot Neural Head Avatars on Mobile Devices
Jianwen Jiang, Gaojie Lin, Zhengkun Rong, Chao Liang, Yongming Zhu, Jiaqi Yang,

Tianyun Zhong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 15920-15929

Existing neural head avatars methods have achieved significant progress in the image quality and motion range of portrait animation. However, these methods prioritize effectiveness over computational overhead. This paper presents MobilePortrait, a lightweight one-shot neural head avatars method that reduces learning complexity by integrating external knowledge into both the motion modeling and image synthesis, enabling real-time inference on mobile devices. Specifically, We introduce a mixed keypoint representation of explicit and implicit keypoints for precise motion modeling and equip it with precomputed visual features to enhance both facial and background synthesis. With these designs, our model can achieve state-of-the-art performance using UNet with low FLOPs as the backbone, requiring less than 1/10 the computational demand. It has been validated to reach speeds of over 50 FPS on mobile devices and support both video and audio-driven inputs.

****

D^3: Scaling Up Deepfake Detection by Learning from Discrepancy

Yongqi Yang, Zhihao Qian, Ye Zhu, Olga Russakovsky, Yu Wu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23850-23859

The boom of Generative AI brings opportunities entangled with risks and concerns. Existing literature emphasizes the generalization capability of deepfake detection on unseen generators, significantly promoting the detector's ability to identify more universal artifacts. This work seeks a step toward a universal deepfake detection system with better generalization and robustness. We do so by first scaling up the existing detection task setup from the one-generator to multiple-generators in training, during which we disclose two challenges presented in prior methodological designs and demonstrate the divergence of detectors' performance. Specifically, we reveal that the current methods tailored for training on one specific generator either struggle to learn comprehensive artifacts from multiple generators or sacrifice their fitting ability for seen generators (i.e., In-Domain (ID) performance) to exchange the generalization for unseen generators (i.e., Out-Of-Domain (OOD) performance). To tackle the above challenges, we propose our Discrepancy Deepfake Detector (D3) framework, whose core idea is to deconstruct the universal artifacts from multiple generators by introducing a parallel network branch that takes a distorted image feature as an extra discrepancy signal and supplements its original counterpart. Extensive scaled-up experiments demonstrate the effectiveness of D3, achieving 5.3% accuracy improvement in the OOD testing compared to the current SOTA methods while maintaining the ID performance. The source code will be updated in our GitHub repository: https://github.com/BigAandSmallq/D3.

****

Jailbreaking the Non-Transferable Barrier via Test-Time Data Disguising

Yongli Xiang, Ziming Hong, Lina Yao, Dadong Wang, Tongliang Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30671-30681

Non-transferable learning (NTL) has been proposed to protect model intellectual property (IP) by creating a "non-transferable barrier" to restrict generalization from authorized to unauthorized domains. Recently, well-designed attack, which restores the unauthorized-domain performance by fine-tuning NTL models on few authorized samples, highlights the security risks of NTL-based applications. However, such attack requires modifying model weights, thus being invalid in the black-box scenario. This raises a critical question: can we trust the security of NTL models deployed as black-box systems? In this work, we reveal the first loophole of black-box NTL models by proposing a novel attack method (dubbed as JailNTL) to jailbreak the non-transferable barrier through test-time data disguising, The main idea of JailNTL is to disguise unauthorized data so it can be identified as authorized by the NTL model, thereby bypassing the non-transferable barrier without modifying the NTL model weights. Specifically, JailNTL encourages unauthorized-domain disguising in two levels, including: (i) data-intrinsic disguising (DID) for eliminating domain discrepancy and preserving class-related content

at the input-level, and (ii) model-guided disguising (MGD) for mitigating output-level statistics difference of the NTL model. Empirically, when attacking state-of-the-art (SOTA) NTL models in the black-box scenario, JailNTL achieves an accuracy increase of up to 55.7% in the unauthorized domain by using only 1% authorized samples, largely exceeding existing SOTA white-box attacks. Code is released at https://github.com/tmllab/2025_CVPR_JailNTL.

****

Light3R-SfM: Towards Feed-forward Structure-from-Motion

Sven Elflein, Qunjie Zhou, Laura Leal-Taixé; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16774-16784

We present Light3R-SfM, a feed-forward, end-to-end learnable framework for efficient large-scale Structure-from-Motion (SfM) from unconstrained image collections. Unlike existing SfM solutions that rely on costly matching and global optimization to achieve accurate 3D reconstructions, Light3R-SfM addresses this limitation through a novel latent global alignment module. This module replaces traditional global optimization with a learnable attention mechanism, effectively capturing multi-view constraints across images for robust and precise camera pose estimation. Light3R-SfM constructs a sparse scene graph via retrieval-score-guided shortest path tree to dramatically reduce memory usage and computational overhead compared to the naive approach. Extensive experiments demonstrate that Light3R-SfM achieves competitive accuracy while significantly reducing runtime, making it ideal for 3D reconstruction tasks in real-world applications with a runtime constraint.

****

Robotic Visual Instruction

Yanbang Li, Ziyang Gong, Haoyang Li, Xiaoqi Huang, Haolan Kang, Guangping Bai, Xianzheng Ma; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 12155-12165

Recently, natural language has been the primary medium for human-robot interaction. However, its inherent lack of spatial precision for robotic control introduces challenges such as ambiguity and verbosity. To address these limitations, we introduce the ***Robotic Visual Instruction (RoVI)***, a novel paradigm to guide robotic tasks through an object-centric, hand-drawn symbolic representation. RoVI effectively encodes spatial-temporal information into human-interpretable visual instructions through 2D sketches, utilizing arrows, circles, colors, and numbers to direct 3D robotic manipulation. To enable robots to understand RoVI better and generate precise actions based on RoVI, we present ***Visual Instruction Embodied Workflow (VIEW)***, a pipeline formulated for RoVI-conditioned policies. This approach leverages Vision-Language Models (VLMs) to interpret RoVI inputs, decode spatial and temporal constraints from 2D pixel space via keypoint extraction, and then transform them into executable 3D action sequences. We additionally curate a specialized dataset of 15K instances to fine-tune small VLMs for edge deployment, enabling them to effectively learn RoVI capabilities. Our approach is rigorously validated across 11 novel tasks in both real and simulated environments, demonstrating significant generalization capability. Notably, VIEW achieves an 87.5% success rate in real-world scenarios involving unseen tasks that feature multi-step actions, with disturbances, and trajectory-following requirements. Code and Datasets in this paper will be released soon.

****

Solving Instance Detection from an Open-World Perspective

Qianqian Shen, Yunhan Zhao, Nahyun Kwon, Jeeeun Kim, Yanan Li, Shu Kong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 9901-9910

Instance detection (InsDet) aims to localize specific object instances within a novel scene imagery based on given visual references. Technically, it requires proposal detection to identify all possible object instances, followed by instance-level matching to pinpoint the ones of interest. Its open-world nature supports its broad applications from robotics to AR/VR but also presents significant challenges: methods must generalize to unknown testing data distributions because (1) the testing scene imagery is unseen during training, and (2) there are domai

n gaps between visual references and detected proposals. Existing methods tackle these challenges by synthesizing diverse training examples or utilizing off-the-shelf foundation models (FMs). However, they only partially capitalize the available open-world information. In contrast, we approach InsDet from an Open-World perspective, introducing our method IDOW. We find that, while pretrained FMs yield high recall in instance detection, they are not specifically optimized for instance-level feature matching. Therefore, we adapt pretrained FMs for improved instance-level matching using open-world data. Our approach incorporates metric learning along with novel data augmentations, which sample distractors as negative examples and synthesize novel-view instances to enrich the visual references. Extensive experiments demonstrate that our method significantly outperforms prior works, achieving >10 AP over previous results on two recently released challenging benchmark datasets in both conventional and novel instance detection settings.

****

Percept, Memory, and Imagine: World Feature Simulating for Open-Domain Unknown Object Detection

Aming Wu, Cheng Deng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4682-4691

To accelerate the safe deployment of object detectors, we focus on reducing the impact of both covariate and semantic shifts. And we consider a realistic yet challenging scenario, namely Open-Domain Unknown Object Detection (ODU-OD), which aims to detect unknown objects in unseen target domains without accessing any auxiliary data. Towards ODU-OD, it is feasible to learn a robust discriminative boundary by synthesizing virtual features. Generally, perception, memory, and imagination are three essential capacities for human beings. Through multi-level perception and rich memory about known objects, the characteristics of unknown objects can be imagined sufficiently, enhancing the ability of discriminating known from unknown objects. Inspired by this idea, an approach of World Feature Simulation (WFS) is proposed, mainly consisting of a multi-level perception, memory recorder, and unknown-feature generator. Specifically, after extracting the features of the input, we separately employ a Mamba and Graph Network to obtain the global-level and connective-level representations. Next, a codebook containing multiple learnable codewords is defined to preserve fragmented memory of known objects. Meanwhile, we perform a modulated operation on the memory to form the imagination bank involving unknown characteristics. Finally, to alleviate the impact of lacking supervision data, based on the multi-level representation and imagination bank, a dedicated unknown-feature generator is designed to recurrently synthesize outlier features deviating from in-distribution (ID) objects. The significant performance gains on four different detection tasks demonstrate the superiorities of our method. The code will be released at https://github.com/AmingWu/WFS.

****

Efficient Depth Estimation for Unstable Stereo Camera Systems on AR Glasses

Yongfan Liu, Hyoukjun Kwon; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6252-6261

Stereo depth estimation is a fundamental component in augmented reality (AR), which requires low latency for real-time processing. However, preprocessing such as rectification and non-ML computations such as cost volume require significant amount of latency exceeding that of an ML model itself, which hinders the real-time processing required by AR. Therefore, we develop alternative approaches to the rectification and cost volume that consider ML acceleration (GPU and NPUs) in recent hardware. For pre-processing, we eliminate it by introducing homography matrix prediction network with a rectification positional encoding (RPE), which delivers both low latency and robustness to unrectified images. For cost volume, we replace it with a group-pointwise convolution-based operator and approximation of cosine similarity based on layernorm and dot product. Based on our approaches, we develop MultiHeadDepth (replacing cost volume) and HomoDepth (MultiHeadDepth + removing pre-processing) models. MultiHeadDepth provides 11.8-30.3% improvements in accuracy and 22.9-25.2% reduction in latency compared to a state-of-t

he-art depth estimation model for AR glasses from industry. HomoDepth, which can directly process unrectified images, reduces the end-to-end latency by 44.5%. We also introduce a multi-task learning method to handle misaligned stereo inputs on HomoDepth, which reduces the AbsRel error by 10.0-24.3%. The overall results demonstrate the efficacy of our approaches, which not only reduce the inference latency but also improve the model performance. Our code is available at https://github.com/UCI-ISA-Lab/MultiHeadDepth-HomoDepth

****

3D-GRAND: A Million-Scale Dataset for 3D-LLMs with Better Grounding and Less Hallucination

Jianing Yang, Xuweiyi Chen, Nikhil Madaan, Madhavan Iyengar, Shengyi Qian, David F. Fouhey, Joyce Chai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29501-29512

The integration of language and 3D perception is crucial for embodied agents and robots that comprehend and interact with the physical world. While large language models (LLMs) have demonstrated impressive language understanding and generation capabilities, their adaptation to 3D environments (3D-LLMs) remains in its early stages. A primary challenge is a lack of large-scale datasets with dense grounding between language and 3D scenes. We introduce 3D-GRAND, a pioneering large-scale dataset comprising 40,087 household scenes paired with 6.2 million densely-grounded scene-language instructions. Our results show that instruction tuning with 3D-GRAND significantly enhances grounding capabilities and reduces hallucinations in 3D-LLMs. As part of our contributions, we propose a comprehensive benchmark 3D-POPE to systematically evaluate hallucination in 3D-LLMs, enabling fair comparisons of models. Our experiments highlight a scaling effect between dataset size and 3D-LLM performance, emphasizing the importance of large-scale 3D-text datasets for embodied AI research. Our results demonstrate early signals for effective sim-to-real transfer, indicating that models trained on large synthetic data can perform well on real-world 3D scans. Through 3D-GRAND and 3D-POPE, we aim to equip the embodied AI community with resources and insights to lead to more reliable and better-grounded 3D-LLMs.

****

LiDAR-RT: Gaussian-based Ray Tracing for Dynamic LiDAR Re-simulation

Chenxu Zhou, Lvchang Fu, Sida Peng, Yunzhi Yan, Zhanhua Zhang, Yong Chen, Jiazhi Xia, Xiaowei Zhou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 1538-1548

This paper targets the challenge of real-time LiDAR re-simulation in dynamic driving scenarios. Recent approaches utilize neural radiance fields combined with the physical modeling of LiDAR sensors to achieve high-fidelity re-simulation results. Unfortunately, these methods face limitations due to high computational demands in large-scale scenes and cannot perform real-time LiDAR rendering. To overcome these constraints, we propose LiDAR-RT, a novel framework that supports real-time, physically accurate LiDAR re-simulation for driving scenes. Our primary contribution is the development of an efficient and effective rendering pipeline, which integrates Gaussian primitives and hardware-accelerated ray tracing technology. Specifically, we model the physical properties of LiDAR sensors using Gaussian primitives with learnable parameters and incorporate scene graphs to handle scene dynamics. Building upon this scene representation, our framework first constructs a bounding volume hierarchy (BVH), then casts rays for each pixel and generates novel LiDAR views through a differentiable rendering algorithm. Importantly, our framework supports realistic rendering with flexible scene editing operations and various sensor configurations. Extensive experiments across multiple public benchmarks demonstrate that our method outperforms state-of-the-art methods in terms of rendering quality and efficiency. Codes will be released to ensure reproducibility.

****

Generative Zero-Shot Composed Image Retrieval

Lan Wang, Wei Ao, Vishnu Naresh Boddeti, Ser-Nam Lim; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 29690-29700

Composed Image Retrieval (CIR) is a vision-language task utilizing queries compr

ising images and textual descriptions to achieve precise image retrieval. This t ask seeks to find images that are visually similar to a reference image while in corporating specific changes or features described textually (visual delta). CIR enables a more flexible and user-specific retrieval by bridging visual data wit h verbal instructions. This paper introduces a novel generative method that augm ents Composed Image Retrieval by Composed Image Generation (CIG) to provide pseu do-target images. CIG utilizes a textual inversion network to map reference imag es into semantic word space, which generates pseudo-target images in combination with textual descriptions. These images serve as additional visual information, significantly improving the accuracy and relevance of retrieved images when int egrated into existing retrieval frameworks. Experiments conducted across multipl e CIR datasets and several baseline methods demonstrate improvements in retrieva l performance, which shows the potential of our approach as an effective add-on for existing composed image retrieval.

****

Large-Scale Text-to-Image Model with Inpainting is a Zero-Shot Subject-Driven Im age Generator

Chaehun Shin, Jooyoung Choi, Heeseung Kim, Sungroh Yoon; Proceedings of the Comp uter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7986-7996

Subject-driven text-to-image generation aims to produce images of a new subject within a desired context by accurately capturing both the visual characteristics of the subject and the semantic content of a text prompt. Traditional methods r ely on time- and resource-intensive fine-tuning for subject alignment, while rec ent zero-shot approaches leverage on-the-fly image prompting, often sacrificing subject alignment. In this paper, we introduce Diptych Prompting, a novel zero-s hot approach that reinterprets as an inpainting task with precise subject alignm ent by leveraging the emergent property of diptych generation in large-scale tex t-to-image models. Diptych Prompting arranges an incomplete diptych with the ref erence image in the left panel, and performs text-conditioned inpainting on the right panel. We further prevent unwanted content leakage by removing the backgro und in the reference image and improve fine-grained details in the generated sub ject by enhancing attention weights between the panels during inpainting. Experi mental results confirm that our approach significantly outperforms zero-shot ima ge prompting methods, resulting in images that are visually preferred by users. Additionally, our method supports not only subject-driven generation but also st ylized image generation and subject-driven image editing, demonstrating versatil ity across diverse image generation applications.

****

MASt3R-SLAM: Real-Time Dense SLAM with 3D Reconstruction Priors

Riku Murai, Eric Dexheimer, Andrew J. Davison; Proceedings of the Computer Visio n and Pattern Recognition Conference (CVPR), 2025, pp. 16695-16705

We present a real-time monocular dense SLAM system designed bottom-up from MASt3 R, a two-view 3D reconstruction and matching prior. Equipped with this strong p rior, our system is robust on in-the-wild video sequences despite making no assu mption on a fixed or parametric camera model beyond a unique camera centre. We introduce efficient methods for pointmap matching, camera tracking and local fus ion, graph construction and loop closure, and second-order global optimisation. With known calibration, a simple modification to the system achieves state-of-th e-art performance across various benchmarks. Altogether, we propose a plug-and- play monocular SLAM system capable of producing globally consistent poses and de nse geometry while operating at 15 FPS.

****

Flow-NeRF: Joint Learning of Geometry, Poses, and Dense Flow within Unified Neur al Representations

Xunzhi Zheng, Dan Xu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 993-1002

Learning accurate scene reconstruction without pose priors in neural radiance fi elds is challenging due to inherent geometric ambiguity. Recent development eith er relies on correspondence priors for regularization or uses off-the-shelf flow estimators to derive analytical poses. However, the potential for jointly learn

ing scene geometry, camera poses, and dense flow within a unified neural represe
ntation remains largely unexplored. In this paper, we present Flow-NeRF, a unifi
ed framework that simultaneously optimizes scene geometry, camera poses, and den
se optical flow all on-the-fly. To enable the learning of dense flow within the
neural radiance field, we design and build a bijective mapping for flow estimati
on, conditioned on pose. To make the scene reconstruction benefit from the flow
estimation, we develop an effective feature enhancement mechanism to pass canoni
cal space features to world space representations, significantly enhancing scene
 geometry. We validate our model across four important tasks, i.e., novel view s
ynthesis, depth estimation, camera pose prediction, and dense optical flow estim
ation, using several datasets. Our approach surpasses previous methods in almost
 all metrics for novel-view view synthesis and depth estimation and yields both
qualitatively sound and quantitatively accurate novel-view flow. Our project pag
e is https://zhengxunzhi.github.io/flownerf/.
****

Viewpoint Rosetta Stone: Unlocking Unpaired Ego-Exo Videos for View-invariant Re
presentation Learning
Mi Luo, Zihui Xue, Alex Dimakis, Kristen Grauman; Proceedings of the Computer Vi
sion and Pattern Recognition Conference (CVPR), 2025, pp. 15802-15812
Egocentric and exocentric perspectives of human action differ significantly, yet
 overcoming this extreme viewpoint gap is critical for applications in augmented
 reality and robotics. We propose ViewpointRosetta, an approach that unlocks lar
ge-scale unpaired ego and exo video data to learn clip-level viewpoint-invariant
 video representations. Our framework introduces (1) a diffusion-based Rosetta S
tone Translator (RST), which, leveraging a moderate amount of synchronized multi
-view videos, serves as a translator in feature space to decipher the alignments
 between unpaired ego and exo data, and (2) a dual encoder that aligns unpaired
data representations through contrastive learning with RST-based synthetic featu
re augmentation and soft alignment. To evaluate the learned features in a standa
rdized setting, we construct a new cross-view benchmark using Ego-Exo4D, coverin
g cross-view retrieval, action recognition, and skill assessment. Our framework
demonstrates superior cross-view understanding compared to previous view-invaria
nt learning and egocentric video representation learning approaches, and opens t
he door to bringing vast amounts of traditional third-person video to bear on th
e more nascent first-person setting.
****

Cross-modal Information Flow in Multimodal Large Language Models
Zhi Zhang, Srishti Yadav, Fengze Han, Ekaterina Shutova; Proceedings of the Comp
uter Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19781-19791
The recent advancements in auto-regressive multimodal large language models (MLL
Ms) have demonstrated promising progress for vision-language tasks. While there
exists a variety of studies investigating the processing of linguistic informati
on within large language models, little is currently known about the inner worki
ng mechanism of MLLMs and how linguistic and visual information interact within
these models. In this study, we aim to fill this gap by examining the informatio
n flow between different modalities---language and vision---in MLLMs, focusing o
n visual question answering. Specifically, given an image-question pair as input
, we investigate where in the model and how the visual and linguistic informatio
n are combined to generate the final prediction. Conducting experiments with a s
eries of models from the LLaVA series, we find that there are two distinct stage
s in the process of integration of the two modalities. In the lower layers, the
model first transfers the more general visual features of the whole image into t
he representations of (linguistic) question tokens. In the middle layers, it onc
e again transfers visual information about specific objects relevant to the ques
tion to the respective token positions of the question. Finally, in the higher l
ayers, the resulting multimodal representation is propagated to the last positio
n of the input sequence for the final prediction. Overall, our findings provide
a new and comprehensive perspective on the spatial and functional aspects of ima
ge and language processing in the MLLMs, thereby facilitating future research in
to multimodal information localization and editing. Our code and collected datas

et are released here: https://github.com/FightingFighting/cross-modal-information-flow-in-MLLM.git.
****

Consistent and Controllable Image Animation with Motion Diffusion Models
Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Tien-Tsin Wong, Yuan-Fang Li, Cunjian Chen; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7288-7298
Diffusion models have achieved significant progress in the task of image animation due to their powerful generative capabilities. However, preserving appearance consistency to the static input image, and avoiding abrupt motion change in the generated animation, remains challenging. In this paper, we introduce Cinemo, a novel image animation approach that aims to achieve better appearance consistency and motion smoothness. The core of Cinemo is to focus on learning the distribution of motion residuals, rather than directly predicting frames as in existing diffusion models. During the inference, we further mitigate the sudden motion changes in the generated video by introducing a novel DCT-based noise refinement strategy. To counteract the over-smoothing of motion, we introduce a dynamics degree control design for better control of the magnitude of motion. Altogether, these strategies enable Cinemo to produce highly consistent, smooth, and motion-controllable results. Extensive experiments compared with several state-of-the-art methods demonstrate the effectiveness and superiority of our proposed approach. In the end, we also demonstrate how our model can be applied for motion transfer or video editing of any given video.
****

Towards Better Alignment: Training Diffusion Models with Reinforcement Learning Against Sparse Rewards
Zijing Hu, Fengda Zhang, Long Chen, Kun Kuang, Jiahui Li, Kaifeng Gao, Jun Xiao, Xin Wang, Wenwu Zhu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 23604-23614
Diffusion models have achieved remarkable success in text-to-image generation. However, their practical applications are hindered by the misalignment between generated images and corresponding text prompts. To tackle this issue, reinforcement learning (RL) has been considered for diffusion model fine-tuning. Yet, RL's effectiveness is limited by the challenge of sparse reward, where feedback is only available at the end of the generation process. This makes it difficult to identify which actions during the denoising process contribute positively to the final generated image, potentially leading to ineffective or unnecessary denoising policies. To this end, this paper presents a novel RL-based framework that addresses the sparse reward problem when training diffusion models. Our framework, named \text B ^2\text -DiffuRL , employs two strategies: **B**ackward progressive training and **B**ranch-based sampling. For one thing, backward progressive training focuses initially on the final timesteps of the denoising process and gradually extends the training interval to earlier timesteps, easing the learning difficulty associated with sparse rewards. For another, we perform branch-based sampling for each training interval. By comparing the samples within the same branch, we can identify how much the policies of the current training interval contribute to the final image, which helps to learn effective policies instead of unnecessary ones. \text B ^2\text -DiffuRL is compatible with existing optimization algorithms. Extensive experiments demonstrate the effectiveness of \text B ^2\text -DiffuRL in improving prompt-image alignment and maintaining diversity in generated images. The code for this work is available.
****

Spatial457: A Diagnostic Benchmark for 6D Spatial Reasoning of Large Mutimodal Models
Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M de Melo, Jieneng Chen, Alan Yuille; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24669-24679
Although large multimodal models (LMMs) have demonstrated remarkable capabilities in visual scene interpretation and reasoning, their capacity for complex and precise 3-dimensional spatial reasoning remains uncertain. Existing benchmarks fo

cus predominantly on 2D spatial understanding and lack a framework to comprehens ively evaluate 6D spatial reasoning across varying complexities.To address this limitation, we present Spatial457, a scalable and unbiased synthetic dataset des igned with 4 key spatial components: multi-object recognition, 2D and 3D spatial relationships, and 3D orientation. Spatial457 supports a cascading evaluation s tructure, offering 7 question types across 5 difficulty levels that progress fro m basic single-object recognition to our newly proposed complex 6D spatial reaso ning tasks.We evaluated various large multimodal models (LMMs) on Spatial457, ob serving a general decline in performance as task complexity increases, particula rly in 3D reasoning and 6D spatial tasks. To quantify these challenges, we intro duce the Relative Performance Dropping Rate (RPDR), highlighting key weaknesses in 3D reasoning capabilities. Leveraging the unbiased attribute design of our da taset, we also uncover prediction biases across different attributes, with simil ar patterns observed in real-world image settings.

****

Omnidirectional Multi-Object Tracking

Kai Luo, Hao Shi, Sheng Wu, Fei Teng, Mengfei Duan, Chang Huang, Yuhang Wang, Ka iwei Wang, Kailun Yang; Proceedings of the Computer Vision and Pattern Recogniti on Conference (CVPR), 2025, pp. 21959-21969

Panoramic imagery, with its 360deg field of view, offers comprehensive informati on to support Multi-Object Tracking (MOT) in capturing spatial and temporal rela tionships of surrounding objects. However, most MOT algorithms are tailored for pinhole images with limited views, impairing their effectiveness in panoramic se ttings. Additionally, panoramic image distortions, such as resolution loss, geom etric deformation, and uneven lighting, hinder direct adaptation of existing MOT methods, leading to significant performance degradation. To address these chall enges, we propose OmniTrack, an omnidirectional MOT framework that incorporates Tracklet Management to introduce temporal cues, FlexiTrack Instances for object localization and association, and the CircularStatE Module to alleviate image an d geometric distortions. This integration enables tracking in panoramic field-of -view scenarios, even under rapid sensor motion. To mitigate the lack of panoram ic MOT datasets, we introduce the QuadTrack dataset--a comprehensive panoramic d ataset collected by a quadruped robot, featuring diverse challenges such as pano ramic fields of view, intense motion, and complex environments. Extensive experi ments on the public JRDB dataset and the newly introduced QuadTrack benchmark de monstrate the state-of-the-art performance of the proposed framework. OmniTrack achieves a HOTA score of 26.92% on JRDB, representing an improvement of 3.43%, a nd further achieves 23.45% on QuadTrack, surpassing the baseline by 6.81%. The e stablished dataset and source code are available at https://github.com/xifen523/ OmniTrack.

****

Potential Field Based Deep Metric Learning

Shubhang Bhatnagar, Narendra Ahuja; Proceedings of the Computer Vision and Patte rn Recognition Conference (CVPR), 2025, pp. 25549-25559

Deep metric learning (DML) involves training a network to learn a semantically m eaningful representation space. Many current approaches mine n-tuples of example s and model interactions within each tuplets. We present a novel, compositional DML model that instead of in tuples, represents the influence of each example (e mbedding) by a continuous potential field, and superposes the fields to obtain t heir combined global potential field. We use attractive/repulsive potential fiel ds to represent interactions among embeddings from images of the same/different classes. Contrary to typical learning methods, where mutual influence of samples is proportional to their distance, we enforce reduction in such influence with distance, leading to a decaying field. We show that such decay helps improve per formance on real world datasets with large intra-class variations and label nois e. Like other proxy-based methods, we also use proxies to succinctly represent s ub-populations of examples. We evaluate our method on three standard DML benchma rks- Cars-196, CUB-200-2011, and SOP datasets where it outperforms state-of-the- art baselines.

****

Enhancing Vision-Language Compositional Understanding with Multimodal Synthetic Data
Haoxin Li, Boyang Li; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24849-24861
Paired image-text data with subtle variations in-between (e.g., people holding surfboards vs. people holding shovels) hold the promise of producing Vision-Language Models with proper compositional understanding. Synthesizing such training data from generative models is a highly coveted prize due to the reduced cost of data collection. However, synthesizing training images for compositional learning presents three challenges: (1) efficiency in generating large quantities of images, (2) text alignment between the generated image and the caption in the exact place of the subtle change, and (3) image fidelity in ensuring sufficient similarity with the original real images in all other places. We propose SPARCL (Synthetic Perturbations for Advancing Robust Compositional Learning), which integrates image feature injection into a fast text-to-image generative model, followed by an image style transfer step, to meet the three challenges. Further, to cope with any residual issues of text alignment, we propose an adaptive margin loss to filter out potentially incorrect synthetic samples and focus the learning on informative hard samples. Evaluation on four compositional understanding benchmarks demonstrates that SPARCL significantly improves the compositionality of CLIP, boosting the average accuracy of the CLIP base model by over 8% across all benchmarks and outperforming state-of-the-art methods by 2% on three benchmarks.
****
Directional Label Diffusion Model for Learning from Noisy Labels
Senyu Hou, Gaoxia Jiang, Jia Zhang, Shangrong Yang, Husheng Guo, Yaqing Guo, Wenjian Wang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25738-25748
In image classification, the label quality of training data critically influences model generalization, especially for deep neural networks (DNNs). Traditionally, learning from noisy labels (LNL) can improve the generalization of DNNs through complex architectures or a series of robust techniques, but its performance improvement is limited by the discriminative paradigm. Unlike traditional ways, we resolve the LNL problems from the perspective of robust label generation, based on diffusion models within the generative paradigm. To expand the diffusion model into a robust classifier that explicitly accommodates more noise knowledge, we propose a Directional Label Diffusion (DLD) model. It disentangles the diffusion process into two paths, i.e., directional diffusion and random diffusion. Specifically, directional diffusion simulates the corruption of true labels into a directed noise distribution, prioritizing the removal of likely noise, whereas random diffusion introduces inherent randomness to support label recovery. This architecture enable DLD to gradually infer labels from an initial random state, interpretably diverging from the specified noise distribution. To adapt the model to diverse noisy environments, we design a low-cost label pre-correction method that automatically supplies more accurate label information to the diffusion model, without requiring manual intervention or additional iterations. Our approach outperforms state-of-the-art methods on both simulated and real-world noisy datasets. Code is available at https://github.com/SenyuHou/DLD.
****
AA-CLIP: Enhancing Zero-Shot Anomaly Detection via Anomaly-Aware CLIP
Wenxin Ma, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, S.Kevin Zhou; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 4744-4754
Anomaly detection (AD) identifies outliers for applications like defect and lesion detection. While CLIP shows promise for zero-shot AD tasks due to its strong generalization capabilities, its inherent Anomaly-Unawareness leads to limited discrimination between normal and abnormal features. To address this problem, we propose Anomaly-Aware CLIP (AA-CLIP), which enhances CLIP's anomaly discrimination ability in both text and visual spaces while preserving its generalization capability. AA-CLIP is achieved through a straightforward yet effective two-stage approach: it first creates anomaly-aware text anchors to differentiate normal an

d abnormal semantics clearly, then aligns patch-level visual features with these anchors for precise anomaly localization. This two-stage strategy, with the help of residual adapters, gradually adapts CLIP in a controlled manner, achieving effective AD while maintaining CLIP's class knowledge. Extensive experiments validate AA-CLIP as a resource-efficient solution for zero-shot AD tasks, achieving state-of-the-art results in industrial and medical applications. The code is available at https://github.com/Mwxinnn/AA-CLIP.
****

HybridGS: Decoupling Transients and Statics with 2D and 3D Gaussian Splatting
Jingyu Lin, Jiaqi Gu, Lubin Fan, Bojian Wu, Yujing Lou, Renjie Chen, Ligang Liu, Jieping Ye; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 788-797
Generating high-quality novel view renderings of 3D Gaussian Splatting (3DGS) in scenes featuring transient objects is challenging. We propose a novel hybrid representation, termed as HybridGS, using 2D Gaussians for transient objects per image and maintaining traditional 3D Gaussians for the whole static scenes. 3DGS is suited for modeling static scenes that assume multi-view consistency, but the transient objects appear occasionally and do not adhere to the assumption, thus we model them as planar objects from a single view by 2D Gaussians. Our novel representation decomposes the scene from the perspective of fundamental viewpoint consistency. Additionally, we present a multi-view supervision method for 3DGS that leverages information from co-visible regions, further enhancing the distinctions between the transients and statics. Then, we propose a straightforward yet effective multi-stage training strategy to ensure robust training and view synthesis. Experiments on benchmarks show our state-of-the-art performance of novel view synthesis in indoor and outdoor scenes, even in the presence of distracting elements. Project page: https://gujiaqivadin.github.io/hybridgs/
****

Keyframe-Guided Creative Video Inpainting
Yuwei Guo, Ceyuan Yang, Anyi Rao, Chenlin Meng, Omer Bar-Tal, Shuangrui Ding, Maneesh Agrawala, Dahua Lin, Bo Dai; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 13009-13020
Video inpainting, which aims to fill missing regions with visually coherent content, has emerged as a crucial technique for creative applications such as editing. While existing approaches achieve visual consistency or text-guided generation, they often struggle to balance coherence and creative diversity. In this work, we introduce VideoRepainter, a two-stage framework that allows users to inpaint a keyframe using established image-level techniques, then propagate the changes to other frames. Our approach can leverage state-of-the-art image models for keyframe manipulation, thereby easing the burden of the video-inpainting process. To this end, we integrate an image-to-video model with a symmetric condition mechanism to address ambiguity caused by direct mask downsampling. We further explore efficient strategies for mask synthesis and parameter tuning to reduce costs in data processing and model training. Evaluations demonstrate our method achieves superior results in both visual fidelity and content diversity compared to existing approaches, providing a practical solution for creative video manipulation.
****

Channel Consistency Prior and Self-Reconstruction Strategy Based Unsupervised Image Deraining
Guanglu Dong, Tianheng Zheng, Yuanzhouhan Cao, Linbo Qing, Chao Ren; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 7469-7479
Recently, deep image deraining models based on paired datasets have made a series of remarkable progress. However, they cannot be well applied in real-world applications due to the difficulty of obtaining real paired datasets and the poor generalization performance. In this paper, we propose a novel Channel Consistency Prior and Self-Reconstruction Strategy Based Unsupervised Image Deraining framework, CSUD, to tackle the aforementioned challenges. During training with unpaired data, CSUD is capable of generating high-quality pseudo clean and rainy image

pairs which are used to enhance the performance of deraining network. Specifica lly, to preserve more image background details while transferring rain streaks f rom rainy images to the unpaired clean images, we propose a novel Channel Consis tency Loss (CCLoss) by introducing the Channel Consistency Prior (CCP) of rain s treaks into training process, thereby ensuring that the generated pseudo rainy i mages closely resemble the real ones. Furthermore, we propose a novel Self-Recon struction (SR) strategy to alleviate the redundant information transfer problem of the generator, further improving the deraining performance and the generaliza tion capability of our method. Extensive experiments on multiple synthetic and r eal-world datasets demonstrate that the deraining performance of CSUD surpasses other state-of-the-art unsupervised methods and CSUD exhibits superior generaliz ation capability. Code is available at https://github.com/GuangluDong0728/CSUD.
****

MobileMamba: Lightweight Multi-Receptive Visual Mamba Network
Haoyang He, Jiangning Zhang, Yuxuan Cai, Hongxu Chen, Xiaobin Hu, Zhenye Gan, Ya biao Wang, Chengjie Wang, Yunsheng Wu, Lei Xie; Proceedings of the Computer Visi on and Pattern Recognition Conference (CVPR), 2025, pp. 4497-4507
Previous research on lightweight models has primarily focused on CNNs and Transf ormer-based designs. CNNs, with their local receptive fields, struggle to captur e long-range dependencies, while Transformers, despite their global modeling cap abilities, are limited by quadratic computational complexity in high-resolution scenarios. Recently, state-space models have gained popularity in the visual dom ain due to their linear computational complexity. Despite their low FLOPs, curre nt lightweight Mamba-based models exhibit suboptimal throughput. In this work, w e propose the MobileMamba framework, which balances efficiency and performance. We design a three-stage network to enhance inference speed significantly. At a f ine-grained level, we introduce the Multi-Receptive Field Feature Interaction (M RFFI) module, comprising the Long-Range Wavelet Transform-Enhanced Mamba (WTE-Ma mba), Efficient Multi-Kernel Depthwise Deconvolution (MK-DeConv), and Eliminate Redundant Identity components. This module integrates multi-receptive field info rmation and enhances high-frequency detail extraction. Additionally, we employ t raining and testing strategies to further improve performance and efficiency. Mo bileMamba achieves up to 83.6% on Top-1, surpassing existing state-of-the-art me thods which is maximum x21| faster than LocalVim on GPU. Extensive experiments o n high-resolution downstream tasks demonstrate that MobileMamba surpasses curren t efficient models, achieving an optimal balance between speed and accuracy.
****

EdgeTAM: On-Device Track Anything Model
Chong Zhou, Chenchen Zhu, Yunyang Xiong, Saksham Suri, Fanyi Xiao, Lemeng Wu, Ra ghuraman Krishnamoorthi, Bo Dai, Chen Change Loy, Vikas Chandra, Bilge Soran; Pr oceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025 , pp. 13832-13842
On top of Segment Anything Model (SAM), SAM 2 further extends its capability fro m image to video inputs through a memory bank mechanism and obtains a remarkable performance compared with previous methods, making it a foundation model for vi deo segmentation task. In this paper, we aim at making SAM 2 much more efficient so that it even runs on mobile devices while maintaining a comparable performan ce. Despite several works optimizing SAM for better efficiency, we find they are not sufficient for SAM 2 because they all focus on compressing the image encode r, while our benchmark shows that the newly introduced memory attention blocks a re also the latency bottleneck. Given this observation, we propose EdgeTAM, whic h leverages a novel 2D Spatial Perceiver to reduce the computational cost. In pa rticular, the proposed 2D Spatial Perceiver encodes the densely stored frame-lev el memories with a lightweight Transformer that contains a fixed set of learnabl e queries. Given that video segmentation is a dense prediction task, we find pre serving the spatial structure of the memories is essential so that the queries a re split into global-level and patch-level groups. We also propose a distillatio n pipeline that further improves the performance without inference overhead. As a result, EdgeTAM achieves 87.7, 70.0, 72.3, and 71.7 J&F on DAVIS 2017, MOSE, S A-V val, and SA-V test, while running at 16 FPS on iPhone 15 Pro Max. The code a

nd models are available at https://github.com/facebookresearch/EdgeTAM.
****
SimLTD: Simple Supervised and Semi-Supervised Long-Tailed Object Detection
Phi Vu Tran; Proceedings of the Computer Vision and Pattern Recognition Conferen ce (CVPR), 2025, pp. 4672-4681
While modern visual recognition systems have made significant advancements, many continue to struggle with the open problem of learning from few exemplars. This paper focuses on the task of object detection in the setting where object class es follow a natural long-tailed distribution. Existing methods for long-tailed d etection resort to external ImageNet labels to augment the low-shot training ins tances. However, such dependency on a large labeled database has limited utility in practical scenarios. We propose a versatile and scalable approach to leverag e optional unlabeled images, which are easy to collect without the burden of hum an annotations. Our SimLTD framework is straightforward and intuitive, and consi sts of three simple steps: (1) pre-training on abundant head classes; (2) transf er learning on scarce tail classes; and (3) fine-tuning on a sampled set of both head and tail classes. Our approach can be viewed as an improved head-to-tail m odel transfer paradigm without the added complexities of meta-learning or knowle dge distillation, as was required in past research. By harnessing supplementary unlabeled images, without extra image labels, SimLTD establishes new record resu lts on the challenging LVIS v1 benchmark across both supervised and semi-supervi sed settings.
****
EarthDial: Turning Multi-sensory Earth Observations to Interactive Dialogues
Sagar Soni, Akshay Dudhane, Hiyam Debary, Mustansar Fiaz, Muhammad Akhtar Munir, Muhammad Sohail Danish, Paolo Fraccaro, Campbell D Watson, Levente J Klein, Fah ad Shahbaz Khan, Salman Khan; Proceedings of the Computer Vision and Pattern Rec ognition Conference (CVPR), 2025, pp. 14303-14313
Automated analysis of vast Earth observation data via interactive Vision-Languag e Models (VLMs) can unlock new opportunities for environmental monitoring, disas ter response, and resource management. Existing generic VLMs do not perform well on Remote Sensing data, while the recent Geo-spatial VLMs remain restricted to a fixed resolution and few sensor modalities. In this paper, we introduce EarthD ial, a conversational assistant specifically designed for Earth Observation (EO) data, transforming complex, multi-sensory Earth observations into interactive, natural language dialogues. EarthDial supports multispectral, multi-temporal, an d multi-resolution imagery, enabling a wide range of remote sensing tasks, inclu ding classification, detection, captioning, question answering, visual reasoning , and visual grounding. To achieve this, we introduce an extensive instruction t uning dataset comprising over 11.11M instruction pairs covering RGB, Synthetic A perture Radar (SAR), and multispectral modalities such as Near-Infrared (NIR) an d infrared. Furthermore, EarthDial handles bi-temporal and multi-temporal sequen ce analysis for applications like change detection. Our extensive experimental r esults on 44 downstream datasets demonstrate that EarthDial outperforms existing generic and domain specific models, achieving better generalization across vari ous EO tasks. Our source codes and pre-trained models are at https://github.com/ hiyamdebary/EarthDial.
****
Learning Endogenous Attention for Incremental Object Detection
Xiang Song, Yuhang He, Jingyuan Li, Qiang Wang, Yihong Gong; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 30354-30364
In this paper, we focus on a challenging Incremental Object Detection (IOD) prob lem. Existing IOD methods follow an image-to-annotation alignment paradigm, whic h attempts to complete the annotations for old categories and subsequently learn s both new and old categories in new tasks. This paradigm inherently introduces missing/redundant/inaccurate annotations of old categories, resulting in a subop timal performance. Instead, we propose a novel annotation-to-instance alignment IOD paradigm and develop a corresponding method named Learning Endogenous Attent ion (LEA). Inspired by the human brain, LEA enables the model to focus on annota ted task-specific objects, while ignoring irrelevant ones, thus solving the anno

tation incomplete problem in IOD. Concretely, our LEA consists of Endogenous Attention Modules (EAMs) and an Energy-based Task Modulator (ETM). During training, we add the dedicated EAMs for each new task and train them to focus on the new categories. During testing, ETM predicts task IDs using energy functions, directing the model to detect task-specific objects. The detection results corresponding to all task IDs are combined as the final output, thereby alleviating the catastrophic forgetting of old knowledge. Extensive experiments on COCO 2017 and Pascal VOC 2007 demonstrate the effectiveness of our method.
****

StarGen: A Spatiotemporal Autoregression Framework with Video Diffusion Model for Scalable and Controllable Scene Generation
Shangjin Zhai, Zhichao Ye, Jialin Liu, Weijian Xie, Jiaqi Hu, Zhen Peng, Hua Xue, Danpeng Chen, Xiaomeng Wang, Lei Yang, Nan Wang, Haomin Liu, Guofeng Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 26822-26833
Recent advances in large reconstruction and generative models have significantly improved scene reconstruction and novel view generation. However, due to compute limitations, each inference with these large models is confined to a small area, making long-range consistent scene generation challenging. To address this, we propose StarGen, a novel framework that employs a pre-trained video diffusion model in an autoregressive manner for long-range scene generation. The generation of each video clip is conditioned on the 3D warping of spatially adjacent images and the temporally overlapping image from previously generated clips, improving spatiotemporal consistency in long-range scene generation with precise pose control. The spatiotemporal condition is compatible with various input conditions, facilitating diverse tasks, including sparse view interpolation, perpetual view generation, and layout-conditioned city generation. Quantitative and qualitative evaluations demonstrate StarGen's superior scalability, fidelity, and pose accuracy compared to state-of-the-art methods.
****

HyperSeg: Hybrid Segmentation Assistant with Fine-grained Visual Perceiver
Cong Wei, Yujie Zhong, Haoxian Tan, Yong Liu, Jie Hu, Dengjie Li, Zheng Zhao, Yujiu Yang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 8931-8941
This paper aims to address universal segmentation for image and video perception with the strong reasoning ability empowered by Visual Large Language Models (VLLMs). Despite significant progress in current unified segmentation methods, limitations in adaptation to both image and video scenarios, as well as the complex reasoning segmentation, make it difficult for them to handle various challenging instructions and achieve an accurate understanding of fine-grained vision-language correlations. We propose HyperSeg, the first VLLM-based universal segmentation model for pixel-level image and video perception, encompassing generic segmentation tasks and more complex reasoning perception tasks requiring powerful reasoning abilities and world knowledge. Besides, to fully leverage the recognition capabilities of VLLMs and the fine-grained visual information, HyperSeg incorporates hybrid entity recognition and fine-grained visual perceiver modules for various segmentation tasks. Combined with the temporal adapter, HyperSeg achieves a comprehensive understanding of temporal information. Experimental results validate the effectiveness of our insights in resolving universal image and video segmentation tasks, including the more complex reasoning perception tasks. Our code is available at https://github.com/congvvc/HyperSeg.
****

Diffusion-based Event Generation for High-Quality Image Deblurring
Xinan Xie, Qing Zhang, Wei-Shi Zheng; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2194-2203
While event-based deblurring have demonstrated impressive results, they are impractical for consumer photos captured by cell phones and digital cameras that are not equipped with the event sensor. To address this problem, we in this paper propose a novel deblurring framework called Event Generation Deblurring (EGDeblurring), which allows to effectively deblur an image by generating event guidance

describing the motion information using a diffusion model. Specifically, we desi gn a motion prior generation diffusion model and a feature extractor to produce prior information beneficial for deblurring, rather than generating the raw even t representation. In order to achieve effective fusion of motion prior informati on with blurry images and produce high-quality results, we develop a regression deblurring network embedded with a dual-attention channel fusion block. Experime nts on multiple datasets demonstrate that our method outperforms state-of-the-ar t image deblurring methods. Our code is available at https://github.com/XinanXie /EGDeblurring https://github.com/XinanXie/EGDeblurring.
****

Video Summarization with Large Language Models
Min Jung Lee, Dayoung Gong, Minsu Cho; Proceedings of the Computer Vision and Pa ttern Recognition Conference (CVPR), 2025, pp. 18981-18991
The exponential increase in video content poses significant challenges in terms of efficient navigation, search, and retrieval, thus requiring advanced video su mmarization techniques. Existing video summarization methods, which heavily rely on visual features and temporal dynamics, often fail to capture the semantics o f video content, resulting in incomplete or incoherent summaries. To tackle the challenge, we propose a new video summarization framework that leverages the cap abilities of recent Large Language Models (LLMs), expecting that the knowledge l earned from massive data enables LLMs to evaluate video frames in a manner that better aligns with diverse semantics and human judgments, effectively addressing the inherent subjectivity in defining keyframes. Our method, dubbed LLM-based Video Summarization (LLMVS), translates video frames into a sequence of captions using a Muti-modal Large Language Model (M-LLM) and then assesses the importanc e of each frame using an LLM, based on the captions in its local context. These local importance scores are refined through a global attention mechanism in the entire context of video captions, ensuring that our summaries effectively reflec t both the details and the overarching narrative. Our experimental results demon strate the superiority of the proposed method over existing ones in standard ben chmarks, highlighting the potential of LLMs in the processing of multimedia cont ent.
****

Sketchtopia: A Dataset and Foundational Agents for Benchmarking Asynchronous Mul timodal Communication with Iconic Feedback
Mohd Hozaifa Khan, Ravi Kiran Sarvadevabhatla; Proceedings of the Computer Visio n and Pattern Recognition Conference (CVPR), 2025, pp. 18176-18186
We introduce Sketchtopia, a large-scale dataset and AI framework designed to exp lore goal-driven, multimodal communication through asynchronous interactions in a Pictionary-inspired setup. Sketchtopia captures natural human interactions, in cluding freehand sketches, open-ended guesses, and iconic feedback gestures, sho wcasing the complex dynamics of cooperative communication under constraints. It features over 20K gameplay sessions from 916 players, capturing 263K sketches, 1 0K erases, 56K guesses and 19.4K iconic feedbacks. We introduce multimodal found ational agents with capabilities for generative sketching, guess generation and asynchronous communication. Our dataset also includes 800 human-agent sessions f or benchmarking the agents. We introduce novel metrics to characterize collabora tive success, responsiveness to feedback and inter-agent asynchronous communicat ion. Sketchtopia pushes the boundaries of multimodal AI, establishing a new benc hmark for studying asynchronous, goal-oriented interactions between humans and A I agents. The dataset can be found at https://sketchtopia25.github.io/
****

Consistency-aware Self-Training for Iterative-based Stereo Matching
Jingyi Zhou, Peng Ye, Haoyu Zhang, Jiakang Yuan, Rao Qiang, Liu YangChenXu, Wu C ailin, Feng Xu, Tao Chen; Proceedings of the Computer Vision and Pattern Recogni tion Conference (CVPR), 2025, pp. 16641-16650
Iterative-based methods have become mainstream in stereo matching due to their h igh performance. However, these methods heavily rely on labeled data and face ch allenges with unlabeled real-world data. To this end, we propose a consistency-a ware self-training framework for iterative-based stereo matching for the first t

ime, leveraging real-world unlabeled data in a teacher-student manner. We first observe that regions with larger errors tend to exhibit more pronounced oscillation characteristics during model prediction.Based on this, we introduce a novel consistency-aware soft filtering module to evaluate the reliability of teacher-predicted pseudo-labels, which consists of a multi-resolution prediction consistency filter and an iterative prediction consistency filter to assess the prediction fluctuations of multiple resolutions and iterative optimization respectively. Further, we introduce a consistency-aware soft-weighted loss to adjust the weight of pseudo-labels accordingly, relieving the error accumulation and performance degradation problem due to incorrect pseudo-labels. Extensive experiments demonstrate that our method can improve the performance of various iterative-based stereo matching approaches in various scenarios. In particular, our method can achieve further enhancements over the current SOTA methods on several benchmark datasets.

****

MV-MATH: Evaluating Multimodal Math Reasoning in Multi-Visual Contexts

Peijie Wang, Zhong-Zhi Li, Fei Yin, Dekang Ran, Cheng-Lin Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 19541-19551

Multimodal Large Language Models (MLLMs) have shown promising capabilities in mathematical reasoning within visual contexts across various datasets. However, most existing multimodal math benchmarks are limited to single-visual contexts, which diverges from the multi-visual scenarios commonly encountered in real-world mathematical applications. To address this gap, we introduce MV-MATH: a meticulously curated dataset of 2,009 high-quality mathematical problems. Each problem integrates multiple images interleaved with text, derived from authentic K-12 scenarios and enriched with detailed annotations. MV-MATH includes multiple-choice, free-form, and multi-step questions, covering 11 subject areas across 3 difficulty levels, and serves as a comprehensive and rigorous benchmark for assessing MLLMs' mathematical reasoning in multi-visual contexts. Through extensive experimentation, we observe that MLLMs encounter substantial challenges in multi-visual math tasks, with a considerable performance gap relative to human capabilities on MV-MATH. Furthermore, we analyze the performance and error patterns of various models, providing insights into MLLMs' mathematical reasoning capabilities within multi-visual settings.

****

Balanced Rate-Distortion Optimization in Learned Image Compression

Yichi Zhang, Zhihao Duan, Yuning Huang, Fengqing Zhu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 2428-2438

Learned image compression (LIC) using deep learning architectures has seen significant advancements, yet standard rate-distortion (R-D) optimization often encounters imbalanced updates due to diverse gradients of the rate and distortion objectives. This imbalance can lead to suboptimal optimization, where one objective dominates, thereby reducing overall compression efficiency. To address this challenge, we reformulate R-D optimization as a multi-objective optimization (MOO) problem and introduce two balanced R-D optimization strategies that adaptively adjust gradient updates to achieve more equitable improvements in both rate and distortion. The first proposed strategy utilizes a coarse-to-fine gradient descent approach along standard R-D optimization trajectories, making it particularly suitable for training LIC models from scratch. The second proposed strategy analytically addresses the reformulated optimization as a quadratic programming problem with an equality constraint, which is ideal for fine-tuning existing models. Experimental results demonstrate that both proposed methods enhance the R-D performance of LIC models, achieving around a 2% BD-Rate reduction with acceptable additional training cost, leading to a more balanced and efficient optimization process. Code will be available at https://gitlab.com/viper-purdue/Balanced-RD.

****

Bridge the Gap: From Weak to Full Supervision for Temporal Action Localization with PseudoFormer

Ziyi Liu, Yangcen Liu; Proceedings of the Computer Vision and Pattern Recognitio

n Conference (CVPR), 2025, pp. 8711-8720
Weakly-supervised Temporal Action Localization (WTAL) has achieved notable succe ss but still suffers from a lack of temporal annotations, leading to a performan ce and framework gap compared with fully-supervised methods. While recent approa ches employ pseudo labels for training, three key challenges: generating high-qu ality pseudo labels, making full use of different priors, and optimizing trainin g methods with noisy labels remain unresolved. Due to these perspectives, we pro pose PseudoFormer, a novel two-branch framework that bridges the gap between wea kly and fully-supervised Temporal Action Localization (TAL). We first introduce RickerFusion, which maps all predicted action proposals to a global shared space to generate pseudo labels with better quality. Subsequently, we leverage both s nippet-level and proposal-level labels with different priors from the weak branc h to train the regression-based model in the full branch. Finally, the uncertain ty mask and iterative refinement mechanism are applied for training with noisy p seudo labels. PseudoFormer achieves state-of-the-art WTAL results on the two com monly used benchmarks, THUMOS14 and ActivityNet1.3. Besides, extensive ablation studies demonstrate the contribution of each component of our method.
****

HomoGen: Enhanced Video Inpainting via Homography Propagation and Diffusion
Ding Ding, Yueming Pan, Ruoyu Feng, Qi Dai, Kai Qiu, Jianmin Bao, Chong Luo, Zhe nzhong Chen; Proceedings of the Computer Vision and Pattern Recognition Conferen ce (CVPR), 2025, pp. 22953-22962
In this paper, we present HomoGen, an enhanced video inpainting method based on homography propagation and diffusion models. HomoGen leverages homography regist ration to propagate contextual pixels as priors for generating missing content i n corrupted videos. Unlike previous flow-based propagation methods, which introd uce local distortions due to point-to-point optical flows, homography-induced ar tifacts are typically global structural distortions that preserve semantic integ rity. To effectively utilize these priors for generation, we employ a video diff usion model that inherently prioritizes semantic information within the priors o ver pixel-level details. A content-adaptive control mechanism is proposed to sca le and inject the priors into intermediate video latents during iterative denois ing. In contrast to existing transformer-based networks that often suffer from a rtifacts within priors, leading to error accumulation and unrealistic results, o ur denoising diffusion network can smooth out artifacts and ensure natural outpu ts. Extensive experiments demonstrate the effectiveness of the proposed method q ualitatively and quantitatively.
****

Generalized Few-shot 3D Point Cloud Segmentation with Vision-Language Model
Zhaochong An, Guolei Sun, Yun Liu, Runjia Li, Junlin Han, Ender Konukoglu, Serge Belongie; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 16997-17007
Generalized few-shot 3D point cloud segmentation (GFS-PCS) adapts models to new classes with few support samples while retaining base class segmentation. Existi ng GFS-PCS methods enhance prototypes via interacting with support or query feat ures but remain limited by sparse knowledge from few-shot samples. Meanwhile, 3D vision-language models (3D VLMs), generalizing across open-world novel classes, contain rich but noisy novel class knowledge. In this work, we introduce a GFS-PCS framework that synergizes dense but noisy pseudo-labels from 3D VLMs with pr ecise yet sparse few-shot samples to maximize the strengths of both, named GFS-V L. Specifically, we present a prototype-guided pseudo-label selection to filter low-quality regions, followed by an adaptive infilling strategy that combines kn owledge from pseudo-label contexts and few-shot samples to adaptively label the filtered, unlabeled areas. Additionally, we design a novel-base mix strategy to embed few-shot samples into training scenes, preserving essential context for im proved novel class learning. Moreover, recognizing the limited diversity in curr ent GFS-PCS benchmarks, we introduce two challenging benchmarks with diverse nov el classes for comprehensive generalization evaluation. Experiments validate the effectiveness of our framework across models and datasets. Our approach and ben chmarks provide a solid foundation for advancing GFS-PCS in the real world. The

code is at here.
****
Do ImageNet-trained Models Learn Shortcuts? The Impact of Frequency Shortcuts on Generalization
Shunxin Wang, Raymond Veldhuis, Nicola Strisciuglio; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25198-25207
Frequency shortcuts refer to specific frequency patterns that models heavily rely on for correct classification. Previous studies have shown that models trained on small image datasets often exploit such shortcuts, potentially impairing their generalization performance. However, existing methods for identifying frequency shortcuts require expensive computations and become impractical for analyzing models trained on large datasets. In this work, we propose the first approach to more efficiently analyze frequency shortcuts at a large scale. We show that both CNN and transformer models learn frequency shortcuts on ImageNet. We also expose that frequency shortcut solutions can yield good performance on out-of-distribution (OOD) test sets which largely retain texture information. However, these shortcuts, mostly aligned with texture patterns, hinder model generalization on rendition-based OOD test sets. These observations suggest that current OOD evaluations often overlook the impact of frequency shortcuts on model generalization. Future benchmarks could thus benefit from explicitly assessing and accounting for these shortcuts to build models that generalize across a broader range of OOD scenarios.
****
HORP: Human-Object Relation Priors Guided HOI Detection
Pei Geng, Jian Yang, Shanshan Zhang; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 25325-25335
Human-Object Interaction (HOI) detection aims to predict the <Human, Interaction, Object> triplets, where the core challenge lies in recognizing the interaction of each human-object pair. Despite recent progress thanks to more advanced model architectures, HOI performance remains unsatisfactory. In this work, we first perform some failure analysis and find that the accuracy of the no-interaction category is extremely low, largely hindering the improvement of overall performance. We further look into the error types and find the mis-classification between no-interaction and with-interaction ones can be handled by human-object relation priors. Specifically, to better distinguish no-interaction from direct interactions, we propose 3D location prior, which indicates the distance between human and object; as of no-interaction vs. indirect interactions, we propose gaze area prior, which denotes whether human can see the object or not. The above two types of human-object relation priors are represented by text and are combined with the original visual features, generating multi-modal cues for interaction recognition.Experimental results on the HICO-DET and V-COCO datasets demonstrate that our proposed human-object relation priors are effective and our method HORP surpasses previous methods under various settings and scenarios. In particular, the usage of our priors significantly enhances the model's recognition ability for the no-interaction category.
****
Building a Mind Palace: Structuring Environment-Grounded Semantic Graphs for Effective Long Video Analysis with LLMs
Zeyi Huang, Yuyang Ji, Xiaofang Wang, Nikhil Mehta, Tong Xiao, Donghyun Lee, Sigmund Vanvalkenburgh, Shengxin Zha, Bolin Lai, Licheng Yu, Ning Zhang, Yong Jae Lee, Miao Liu; Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24169-24179
Long-form video understanding with Large Vision Language Models is challenged by the need to analyze temporally dispersed yet spatially concentrated key moments within limited context windows. In this work, we introduce VideoMindPalace, a new framework inspired by the "Mind Palace", which organizes critical video moments into a topologically structured semantic graph. VideoMindPalace organizes key information through (i) hand-object tracking and interaction, (ii) clustered activity zones representing specific areas of recurring activities, and (iii) environment layout mapping, allowing natural language parsing by LLMs to provide gro

unded insights on spatio-temporal and 3D context. In addition, we propose the Video MindPalace Benchmark (VMB), to assess human-like reasoning, including spatial localization, temporal reasoning, and layout-aware sequential understanding. Evaluated on VMB and established video QA datasets, including EgoSchema, NExT-QA, IntentQA, and the Active Memories Benchmark, VideoMindPalace demonstrates notable gains in spatio-temporal coherence and human-aligned reasoning, advancing long-form video analysis capabilities in VLMs.
****